Supporting Professional Capacity of Teachers Through Teacher Evaluation: Associations of School Climate, Teacher Evaluation, and Professional Capacity in Four Countries

A DISSERTATION

SUBMITTED TO THE FACULTY OF

THE UNIVERSITY OF MINNESOTA

BY

Jisu Ryu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTORATE OF PHILOSOPHY

Dr. Karen Seashore, Advisor

December 2020

Acknowledgements

I am grateful for so many individuals who supported and encouraged me throughout the long journey to develop and write this dissertation.

First, I would like to thank my advisor, Dr. Karen Seashore, who grounded and deepened my understanding of educational research with keen insights and a caring heart. You deeply understood my passion as a researcher and my life as a scholar, woman, and mother of two children. I am also grateful for my committee members for their thoughtful advice and guidance. I thank Dr. Peter Demerath, who guided me throughout my coursework. Thanks to your rich knowledge of philosophies and theories, I could be a more thoughtful researcher who attempts to see beyond a phenomenon. Thank you, Dr. Nicola Alexander, for the conversations we had and your teaching in policy and politics that grounded my current work formulating and developing educational policies in the field. I also appreciate Dr. Ernest Davenport, who strengthened my knowledge in quantitative methods and provided insightful advice and helpful guidance.

Thank you to Dr. Youngju Ryu, who is my role model and my beloved cousin. You shed light on my career path in the U.S., which changed my life.

I thank many educational leaders, researchers, and practitioners for the time that they shared with me and their insights that guided me throughout. I learned so much from them.

I am grateful to my EPL peers: Clare Halloran, Hyunjun Kim, Jeff Walls, Malai Turnbull, Sammy Holoquist, Sara Kemper, and Sung Tae Jang. I was very fortunate to

have friends like you who are fun, encouraging, and supportive. I am so proud that we all have completed the Ph.D. race.

Thank you to my friends for your long-time friendship and support. Hyunae, Hyunoh, Jung min, Sun-Young, and Yeryoung, you see me and support me as I am.

Thank you to Jane and Raewon's family for your continuous and loving support and fellowship. Your family is a precious gift in our Seattle life and beyond.

Thank you to my colleagues at the Professional Educator Standard Board. It is a rare opportunity to work with competent and like-minded colleagues who have a genuine passion for education.

To my family, I appreciate your unconditional support and love. In particular, Mom and Dad, you have provided me with countless encouragement and showed endless confidence in me, which helped set out my path and life. I also truly appreciate my parents-in-law for their consistent prayer and encouragement throughout this journey. You showed me diligence and wisdom. My sisters, Jooyoun and Nasun, my pride, my three Musketeers, thank you for supporting my decisions even in the most challenging times and for being sisters whom I can always count on. Thank you to my brother-in-law, Joongho Park, and my precious nephew and niece, Seojoon and Seoyoon, for your support and love that always make me smile. Thank you, aunt Young-soon Ryu, for your gentle, kind, and loving support.

Lastly, my enormous gratitude to my husband and friend, Minho, and our kids Liam Joonwoo and Katelyn Siwoo who walked with me throughout this long journey. Minho, I want to thank you from my deepest heart for all you have done to help me

complete my dissertation. Especially, thank you for taking care of the kids on weekends so that I could work on my research—taking them to parks, playing sports and games together, and making them laugh. It is so precious that we are building our future together in faith and love. We did it together. Joonwoo and Siwoo, thank you for understanding and cheering for Mom. You two and your futures have constantly motivated me to complete my Ph.D. Thank you for coming to us and being our family. I look forward to spending more time with you and playing your favorite board games.

Abstract

In many countries, teacher evaluation has been viewed as a policy lever to improve the quality of teaching and student achievement. Recent research suggests that teacher evaluation can also be implemented as a mechanism for professional growth with careful consideration of the organizational context. However, few studies have examined the way in which a teacher evaluation policy may result in the improvement of teaching.

Two key features of effective teacher evaluation are first, balancing the two purposes of accountability and professional growth and second, implementing teacher evaluation policies with a shared responsibility among teachers, administrators, and government agencies. This dissertation explored the applicability of this analytical framework within schools by examining associations among teacher evaluation outcomes, their impact on various aspects of teaching, and a school climate of shared responsibility. The framework was used to examine teacher evaluation in four countries: the U.S., Finland, South Korea, and Japan.

This study delved into the following questions: *(1) How are national teacher evaluation policies implemented at the local level in four countries? (2) How are teacher evaluation policies and school climate associated with teachers' perceived professional capacity in four countries? (3) How are teacher evaluations associated with teachers' professional capacities when evaluation is accompanied by support of teacher professional growth?* These research questions were investigated using the Organization for Economic Cooperation and Development (OECD) 2013 surveys of principals and teachers.

The findings showed that, despite differences in national policies to reform teacher evaluation, teachers in all four countries still viewed teacher evaluation largely as an administrative requirement. However, further analysis revealed that teachers were more likely to perceive that teacher evaluation was positively associated with their professional capacities when it was coupled with school-level actions to support their professional growth and a school-level climate of shared responsibility. Because teacher evaluation policies exist in most countries, the potential impact of improved implementation at local and national levels could be substantial.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

Improving teacher quality is at the center of global policy debates and reform agendas in many countries. Notably, many of these global policies have been developed under strikingly similar policy assumptions: holding teachers accountable for their teaching will enhance the quality of the teaching workforce (Akiba, 2017; Paine & Zeichner, 2012). Since the early 1990s, the U.S. federal government has chosen to prioritize accountability over professional development by increasing regulatory standardized monitoring (Cohen-Vogel, 2005; Plecki & Loeb, 2004). Policies have centered on teacher evaluation as a policy lever to uphold teacher responsibility for student achievement. However, less consideration has been given to the educative function of teacher evaluation: providing ample feedback and opportunities for reflection that stimulates self-directed improvement.

Indeed, a growing body of research shows that teacher evaluation, when designed and used thoughtfully, can balance these dual purposes of accountability and reflective practice and, thus, can be an effective tool to enhance classroom practice (Hargreaves & Braun, 2013; Murphy, Hallinger, & Heck, 2013; Smylie, 2014). Given these policy and research contexts, this study asks, "what are effective teacher evaluation policies that balance both accountability and improvement, and how can they be successfully implemented?" Using a well-established international survey data, the 2013 Teaching and Learning International Survey (TALIS) sponsored by the Organization for Economic Cooperation and Development (OECD), allows an examination of this question in both the U.S. and other countries. This research topic is timely, as the Every Student Succeeds

Act (ESSA) encouraged more flexibility for teacher evaluation (National Education Association, 2015; Sawchuk, 2016; Taylor & Tyler, 2012b), providing opportunities to rethink effective teacher evaluation policy.

Any study addressing teacher evaluation is fundamentally rooted in questions of who a good teacher is and what is good teaching. Teacher quality is traditionally viewed as an individual trait of teaching practice, but there has been an increased emphasis on the influence of school and district contexts on teaching practice in recent studies (Kennedy, 2008; Knight et al., 2015). This line of research suggests the limits to the assumption that improving individual teachers' knowledge and skills will result in increased learning. It also indicates a need to examine how local policies, interactions among teachers, and teacher-administrator relationships influence teachers' classroom practice. Thus, an organizational approach begins with the assumption that educational system change can support quality teaching.

From this viewpoint, two features of teacher evaluation have emerged as key elements: (1) balancing both purposes of accountability and professional growth and (2) developing shared responsibility in schools (Darling-Hammond, 2013; Hargreaves & Braun, 2013). Notably, each of these emphasizes the importance of social interactions among actors in schools. The idea of using evaluation as a feedback tool for professional growth and the concept of shared responsibility is relatively newer topics in education policy discourse that has been focused on the high-stakes nature of teacher evaluation. However, they have recently drawn attention from scholars who are interested in both teaching quality and quality of support for teachers. For example, early studies indicated

that local contexts and shared accountability affected teachers' responses to state and district evaluation policies (Ingram, Louis, & Schroeder, 2004; Louis, Febey, & Schroeder, 2005). Based on an extensive survey, Hargraves and Braun (2015) argue that these two features are critical elements of effective teacher evaluation.

This study attempts to advance this notion of teacher evaluation research by focusing on the implementation of teacher evaluation policies in multiple countries. The following research questions guide the investigation: *(1) How are national teacher evaluation policies implemented at the local level in four countries? (2) How are teacher evaluation policies and school climate associated with teachers' perceived professional capacity in four countries? (3) How are teacher evaluations associated with teachers' professional capacities when evaluation is accompanied by support of teacher professional growth?* These questions are nested in schools, where the implementation of many national/state/district teacher policies becomes entwined with school-level policies and practices (Honig, 2006; Spillane, Reiser, & Reimer, 2002).

This study uses data collected by the Organizational for Economic Cooperation and Development (OECD) to examine the relationships among critical factors of teacher evaluation and teaching practice across national education systems. The "core" target population for TALIS 2013 (Teaching and Learning International Survey) was teachers and school leaders in lower secondary schools in 34 nations (OECD, 2014). (More details on the survey will be presented in Chapter 3).

I have selected four countries for comparative analysis: The U.S., Japan, South Korea, and Finland. The non-US countries are consistently ranked above the international

average in reading, math, and science Program for International Student Assessment (PISA), and are known for their stable teacher hiring and support systems. Thus, understanding the implementation of teacher evaluation policies in these three nations will provide valuable insights for the educators and policymakers in the U.S. in which efforts to experiment with teacher evaluation have emerged more recently.

This study focuses on comparing the patterns of relationships among variables related to teacher evaluation across the four nations. The 2013 TALIS teacher survey provides ample data on teachers' experiences and could be linked to the principal survey, which provides data on standard school policies and practices related to teacher evaluation. The descriptive statistics of principals were analyzed to capture school formal teacher evaluation policy in each country. The teacher survey was examined to understand how teacher evaluation was implemented at the school level, including (1) teacher evaluation outcomes that could support professional growth; (2) a climate of shared responsibility within the school; and (3) positive relationships with evaluation on teachers' professional capacities.

Much of the research on teacher evaluation has focused on the validity of evaluation measures, but fewer studies have analyzed how teacher evaluation policies are implemented. An examination of the implementation of teacher evaluation will provide meaningful insight on effective strategies for leaders who hope to improve teacher evaluation in their schools. Further, current studies of teacher evaluation policy that focus on teachers' professional growth include robust qualitative case studies; by using a quantitative approach, this study broadens the scope of research on this topic. Finally,

analyzing cross-national similarities and differences in how teacher evaluation policies are implemented can deepen our understanding of how different approaches to teacher evaluation may affect other initiatives to improve teacher quality.

Chapter 2 Literature Review

Global and Local Teacher Evaluation Policy Contexts

Improving teacher performance has been a global policy concern for several decades. An increasing number of countries has developed and implemented a wide range of teacher policies, from teacher recruitment, development, and retention, to evaluation with the goals of improving student performance and teaching quality (Akiba, 2017; Darling-Hammond et al., 2017; OECD, 2005). Researchers show that the global policy discussion regarding teacher and teaching quality has developed under noticeably similar policy rhetoric and assumptions in different nations (Akiba, 2017; Paine & Zeichner, 2012). First, many countries have determined that "improving teacher quality is critical for educating future citizens who are equipped with global competitiveness in the new world of knowledge economy" (Akiba, 2017, p. 155). Second, it is generally believed that holding individual teachers accountable will enhance the quality of the teaching workforce.

Consideration of these global contexts may seem far from the reality of domestic, local-level policies. However, Akiba (2017) keenly pointed out that, both at the national and local levels, policy debates around a "teacher quality" problem and the solutions identified by policy actors were influenced by global dynamics. Interactions among global policymakers have been more active than ever before, given the development of international assessments and ranking system reports from international organizations like OECD, UNESCO, and World Bank, as well as international meetings and programs. Through these international networks, policymakers have shared "a sense of urgency

among many countries for implementing a teacher reform for improving student achievement" (Akiba, 2017, p.157).

The U.S. federal government's efforts to devise a rigorous teacher evaluation system across states corresponded to the pressure to develop a neo-liberal model of global competitiveness (Cohen-Vogel, 2005) and widely influenced local-level policies. Simultaneously, within this "global convergence" of teacher-related reforms, cross-national differences and divergence were emerged as a result of "collective sensemaking, negotiation and contestation within nation-specific teaching and policy environments" (Akiba, 2017, p.162). Accordingly, comprehending how global dynamics have influenced the implementation of teacher evaluation policies and actual classroom teaching will deepen the understanding of policymakers and researchers.

This paper examines teacher evaluation policy in the United States, its associations with the local educational contexts, and its impact on teachers, along with three countries with high-performing education systems: Finland, South Korea, and Japan. Broadly, their education systems have been developed in different cultural backgrounds and, structurally, the ways in which their teacher evaluations have been evolved were quite different. It provides a fertile ground for a comparative study. Moreover, these countries initiated national teacher evaluation policy reforms almost at the same time as the U.S., taking into account the global dynamics. However, each country overhauled its national teacher evaluation policy differently depending on their teaching and policy environments. In addition, the non-U.S. countries are consistently ranked above the international average in reading, math, and science PISA and are known

for their stable teacher hiring and support systems. Examining the practices in high-performing nations can provide meaningful insights on how teacher evaluation policy should be implemented elsewhere. Please note that other high-performing nations, such as China and Canada, were not selected because their samples were drawn from certain regions of each nation and did not necessarily represent the sample nation-wide. A country like Singapore, which is a city-state, was similarly not chosen as representing a national sample.

The following sections provide an overview of policy contexts and reform agendas in each country. While teacher evaluation was used mainly to improve teacher accountability using a high-stakes structure, the comparative policy analysis shows that high accountability does not necessarily require high-stakes outcomes.

United States

Education in the United States is best characterized as a decentralized and loosely coupled system where local agencies have strong control over school-level decisions. Some view it as a democratic system, while others perceive it as fragmented (Cohen, 2010; Darling-Hammond et al., 2017; Labaree, 2012). Since the early 1990s, the U.S. federal government has strengthened the role of federal policy that regulated and monitored teacher quality, choosing to prioritize accountability over professional development in its approach to teacher evaluation (Cohen-Vogel, 2005; Plecki & Loeb, 2004). This policy trend corresponded to the pressure to develop a neo-liberal model of global competitiveness and was consistent with the standardized reform initiatives that have emerged within the U.S. since the 1980s (Akiba, 2017; Demerath, Lynch, &

Davidson, 2008; Lipman, 2013). Various teacher policies have been implemented under the catchphrase that improving teacher quality was key to enhancing student learning, national competitiveness in the global knowledge economy, and meeting society's expectations of social justice and equity (Akiba, 2017; Berry, Darling-Hammond, Hirsch, Robinson, & Wise, 2006; Cochran-Smith, Piazza, & Power, 2013; Hallgren, James-Burdumy, & Perez-Johnson, 2014). Policymakers at the national level conceived of plans to hold educators accountable for student performance, as exemplified in initiatives such as the Highly Qualified Teacher (HQT) mandate in the No Child Left Behind (NCLB) Act of 2001 and funding for the Race to the Top (RTT) in 2009.

These federal regulations initiated two significant shifts in teacher evaluation in schools. First, they increased individual teacher accountability for student learning by matching student achievement information with a particular teacher's evaluation. Second, federal regulations began to address and specify ways to evaluate teachers to improve student learning outcomes (Cochran-Smith et al., 2013; Harris & Herrington, 2015; Pennington & Mead, 2016). During this process, alternative evaluation measures that could gauge teacher effectiveness, including Value-Added Models (VAMs) or student percentile rates, emerged in federal and state policy agendas (Hull, 2013). States and school districts used these evaluation models to make key personnel decisions about teacher retention, dismissal, and compensation.

However, these high-stakes teacher evaluation policy initiatives have inspired fierce debates in research and policy circles. Many have questioned the validity of available "objective measures" as a gauge of teacher quality and accountability (i.e., the

extent to which they genuinely evaluate teaching practice and teacher performance) and were concerned about the ways in which effectiveness measures, such as VAMs, have influenced teacher personnel decisions (Hargreaves & Braun, 2013). Furthermore, Superfine, Gottlieb, and Smylie (2012) argued that, while RTT promoted teacher accountability, it was silent on teacher development actions that might address identified weaknesses. As the critiques of and unintended consequences from the high-stakes federal regulations accrued, the Obama administration enacted the Every Student Succeeds Act (ESSA) in 2015. Most notably, ESSA loosened the reins on accountability in the education sector, which were linked tightly to student achievement scores, and granted more flexibility to state and local governments (National Education Association, 2015; Sawchuk, 2016). The federal government thereby took on a more development-focused initiative to improve teaching quality by pushing the field of teacher assessment forward through innovation.

South Korea

South Korea (Korea, hereafter referred to as Korea) has developed a strong and highly qualified teaching workforce (Darling-Hammond et al., 2017). Teaching is one of the popular career choices in present-day Korea, with jobs in the education sector providing high social status, job stability, and stable pay. In Korea, most public school teachers are employed with tenure as public civil servants (Kim & Youngs, 2016). According to the Center on the International Education Benchmarks, or CIEB, (2020), only 5% of applicants are accepted into primary school teacher training programs, and the

proportion of Korean teachers who are fully certified and hold bachelor's degrees is among the highest in the world.

The traditional teacher evaluation system, introduced in 1964, was based upon teacher performance ratings and incorporated these ratings when determining teacher promotion and school rotation (Choi & Park, 2016). However, the traditional performance rating system faced criticism because it was used mainly for promotions to education specialists or principals but did not consider teaching practices when assessing teacher rating criteria. Thus, this system had little impact on improving teaching quality. As the concerns regarding the effectiveness of the teacher evaluation policy escalated and international interest in teacher quality increased, Korean policymakers implemented two separate teacher evaluation reforms: a performance-based pay system and, several years later, a professional development system.

In 2001, the teacher performance-based pay system was first introduced to generate competition among teachers and offer a financial bonus. It was first discussed in Korea following the severe economic crisis of the late 1990s. Through this initiative, policymakers aimed to promote a performance-centered work environment and constructive competition using financial rewards for teachers' efforts (Choi & Park, 2016). Although the Ministry of Education set out guidelines with examples, each school was responsible for determining its respective evaluation criteria. Policymakers introduced this system in 2001; however, it was not rolled out nationally until 2005, as it faced backlash and strong opposition from teachers (Ha & Sung, 2011; Yoo, 2018). A majority of teachers expressed concerns about a lack of consensus and trust in the

performance-based system's evaluation criteria. Furthermore, this system's evaluation criteria did not address teaching practices, and thus teachers felt that the criteria were irrelevant to what they taught in their classrooms (Seo, 2012). As a result, many teachers perceived that the outcomes of this teacher evaluation were not meaningful and did not appreciate the monetary rewards that they received. In an interview featured in Ha and Sung's study (2011), many teachers asserted that earning a bonus was not a primary motivator to improve their teaching practices. Some even felt it was an insult to their professional dedication to teaching.

In 2010, a second teacher evaluation system, which focused on professional development this time, was established nationwide to provide teachers with formative feedback on their teaching practices and support developing their professional competencies (Choi & Park, 2016). In 2004, the OECD report analyzed several critical issues in the existing Korean teacher evaluation system (Coolahan, Santiago, Phair, & Ninomiya, 2004). They pointed out that the performance rating for promotion was lack of clear and systemic evaluation standards and procedures and, thus, failed to provide constructive learning opportunities. Simultaneously, there was a growing concern around the ineffectiveness of the traditional teacher evaluation system among researchers and education leaders in Korea (Choi & Park, 2016). Responding to the global and national concerns around the effectiveness of the teacher evaluation system, the Ministry of Education and Human Resource Development[1] developed a new form of teacher evaluation in 2005, which was the teacher evaluation system for professional

[1] The Ministry of Education changed its name to the Ministry of Education and Human Resources in 2001.

development. In 2010, the teacher evaluation for professional development system was established nationwide. It purported to develop the teaching skills and competencies of teachers by providing feedback and customized training programs. The new teacher evaluation used multiple measures and multiple evaluators, including peer review by three colleague teachers, evaluation from school principals, student surveys for grades 4-12, and parent surveys (Seo, 2012). Teachers who received low scores were required to take 60 hours of professional development training, and teachers who received high scores could take 6-12 months of sabbatical for research.

Views of the teacher evaluation system for professional development were mixed (Kang, 2013; Kim & Youngs, 2016; Kim & Kim, 2012; Seo, 2012; Yoo, 2018). Students and parents believed that the new teacher evaluation policy led teachers to put more effort and time into teaching. Teachers, on the other hand, reported that it had little impact on their professional learning and the improvement of their teaching practices. According to Yoo (2018), teachers opposed the new evaluation policy because of its intense focus on requirements rather than professional support, a lack of consensus on the evaluation criteria, and the unreliable source of evaluation. Research showed that, even though the new model purportedly aimed to promote professional development, teachers perceived it as emphasizing teacher accountability in practice. For example, teachers considered the training given to teachers receiving low scores as a punishment rather than constructive support. Kim and Youngs (2016) argued that this misalignment between policymakers' expectations and teachers' implementation behavior negatively impacted policy implementation and its results.

Even though the teacher evaluation system for professional development was criticized by educators, Seo (2012) and Yoo (2018) determined that it was successfully implemented when there were consensus and collaboration among teachers and school leaders. For example, at one school where a collaborative team observed classroom instruction and identified teachers' strengths and weaknesses, there was an improvement in student achievement and school-level evaluation.

Japan

In Japan, the teaching profession has been known for being highly selective, especially at the hiring stage (CIEB, 2020). To become a teacher in Japan, candidates go through a rigorous set of school board exams and evaluations. However, once Japanese teachers were hired, the teacher evaluations that they received were for relatively low stakes and were largely done as an administrative task. Thus, there have been growing concerns that the traditional performance rating system was not effective in improving teacher morale and performance (Aspinall, 2001).

In the early 2000s, policymakers in Japan initiated teacher evaluation reforms as a way to strengthen each teacher's accountability for student learning (Katsuno, 2016). At the same time, Japanese policymakers expressed their concerns about the country's competitiveness in global markets, highlighting that Japan should be able to compete with rapidly growing countries. In 2006, the Council on Economic and Fiscal Policy noted that "Human resources who are well qualified to engage in international activities and who are also going to be the main players in the future labour market must be secured…For this purpose, we will aim to achieve world top-level performance in

international academic ability tests by 2010" (Shimonkaigi, 2006, as cited in Katsuno, 2016, p. 23). A growing number of people have also shared this concern regarding the effectiveness of the education system and have advocated for linking teacher evaluation results with teacher compensation.

As a reaction, many local prefecture boards of education have implemented a new teacher evaluation model in which teachers are required to participate in an annual cycle of evaluation. In this cycle, teachers set annual goals, discuss their goals with head teachers, conduct self-assessment, and are formally evaluated by head teachers. The government promoted this as a professional development model that would support teachers in achieving their performance goals and would encourage collaboration between the teachers and head teachers. A survey conducted by the Ministry of Education, Culture, Sports, Science, and Technology in 2010 showed that this new model of teacher evaluation had been widely implemented across the nation. Some regions have even attempted to link the new teacher performance evaluation scores with financial bonuses that teachers receive.

To understand the impact of this teacher evaluation reform, Katsuno (2016) analyzed policy narratives and teacher surveys conducted in Tokyo and Osaka. Based on his findings, he argued that the new teacher evaluation did not seem to contribute to professional development as was intended. Rather, it drove changes in the power relationships in schools by positioning head teachers as formal evaluators, emphasized teacher accountability using the new teacher evaluation model, and resulted in top-down management. Moreover, teachers expressed that they felt pressured to compete with their

colleagues and had fewer opportunities to act as a collective voice in the decision-making process in their schools. It seemed that the new model of teacher evaluation increased competition, demoralized collaboration among teachers, and infused bureaucracy into the system.

Finland

Finland is touted for its high student achievement and excellent teacher quality (CIEB, 2020; Darling-Hammond et al., 2017). The quality of teachers is often cited as a key factor in the success of the Finnish education system. The teaching profession is one of the most admired careers in Finland mainly because of the high standards in its selection process, the work itself, and the working conditions, rather than its teacher salaries. In Finland, there has been a systemic effort to promote teacher quality that was initiated by reform movements from the top; for example, every teacher is required to earn a master's degree for permanent employment, and the coursework and clinic program of teacher education programs have been strengthened (Sahlberg, 2011a). This has made the teaching profession more highly selective and intensive enough to increase teaching capacities to the degree that teachers are respected as professionals like doctors and lawyers.

Given the rigorous teacher education system in Finland, different types of systemic support and flexibility are provided once individuals are employed as teachers. Finnish teachers expect that they will be given a full range of professional autonomy to practice what they have been educated to do. As in Korea, Finnish teachers also work as public servants for specific municipalities, which assures their job security (Tarhan,

Karaman, Lauri, & Aerila, 2019). In Finland, teachers value their jobs because of the autonomy, collaboration, and research and development opportunities in the teaching profession and the trust people have in teachers and their professional judgments (Darling-Hammond et al., 2017; Sahlberg, 2011a).

Finland, like many other countries selected for this study, underwent a massive teacher evaluation reform, but its approach was the opposite of most typical efforts. Beginning in the 1990s, local authorities were given more autonomy from the state administration; a formerly strict national inspection system became, through this process, a more locally developed monitoring system (OECD, 2011; Tarhan et al., 2019; Webb* et al., 2004). By 2000, the Finnish Ministry of Education eliminated all official inspection mechanisms of the work of teachers, such as inspector visits, the state-mandated curriculum, compulsory use of certain materials, rigid teacher schedules, and class journals (Tarhan et al., 2019). Finland also forewent a nationally regulated framework of teacher evaluation and standardized tests measuring student achievement, meaning there was no formal consideration of student learning outcomes connected to teacher evaluations (Sahlberg, 2011b). Instead, teachers were evaluated based on the progress on a self-developed plan for improvement in their subject area of teaching. Teachers also received feedback from their principal and the school staff. The purpose of the teacher evaluation in Finland clearly emphasized teacher empowerment.

Finnish education policy also was influenced by the global dynamics that emphasized teacher accountability, yet the ways in which it influenced policy narratives in Finland looked different from most other countries. Unlike countries where the global

dynamics of international testing and the emphasis on teaching quality initiated massive teacher evaluation reforms, in Finland, these ensured that their efforts in reforming the education system were successful. High achievement in PISA and the increased global reputation of their education system have reinforced Finnish educators' belief in the value of their teacher evaluation reform. Since the 1970s, a great amount of progress has been made in changing the way teachers are educated, and they have become highly trusted professionals. This, in return, has enhanced teacher autonomy and trust in their professional judgment. Finnish educators regarded that their teachers' accountability was demonstrated through teacher autonomy and shared responsibility. As Sahlberg (2011a) described, "the shared responsibility for teaching and learning characterizes how educational accountability is arranged in Finland" (p. 130). This different approach to accountability offered Finnish teachers a strong sense of professional responsibility and initiative.

Theoretical Perspectives

The debates around what constitutes a good teacher and how to evaluate teacher quality are essentially political and ideological (Pennington & Mead, 2016; Sato, 2014). New policy ideas and agendas compete and are selected during contentious debates in the policy and research circles (Cochran-Smith et al., 2013). In this agenda-setting process, prominent researchers and leading policymakers play significant roles in creating policy discourses and leveraging political power (Alexander, 2012; Kingdon, 1984). Likewise, teacher evaluation policies do not arise in a vacuum. The historical contexts for teacher evaluation policies also suggest that different policies have been selected, marginalized,

and silenced through fierce and deliberate debates. In the research circle, scholars define teacher quality differently depending on their epistemological and theoretical grounds and therefore suggest different teacher evaluation policy agendas. In this regard, Cochran-Smith, Piazza, & Power (2013) keenly point out that policies are developed formally and informally on multiple levels and in multiple forms, and that policy discourses are both discrete and interconnected.

Regarding teacher evaluation policy, theorists have used three distinct disciplinary perspectives to provide meaningful and critical insights about teacher evaluation over the past few decades. They include (1) economists who draw on human capital/human resource theories, (2) teacher-practice theorists who delve into issues about teacher education and professional growth, and (3) organizational theorists who analyze the teaching workforce from organizational and ecological perspectives. Comparing these theoretical disciplines can deepen and broaden our understanding of teaching quality and evaluations thereof. While these three groups have often engaged in fierce debates about teacher evaluation policies, they also have overlapping key assumptions about them. In this sense, this review discusses not only how these theories have developed different conceptualizations of teacher evaluation policies but also how their findings can complement one another.

Economics and Human Resource Theories

Scholars in economics and human resources, who adopt a rationalist perspective, have long investigated teacher evaluation instruments that are accurate, fair, generalizable, efficient, and feasible. They have argued that the traditional U.S. education

system is problematic because it treats most teachers similarly, although their effects on student achievement vary widely (Hanushek & Rivkin, 2006; Pennington & Mead, 2016; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Despite their varied influence on student learning, most teachers receive similar ratings, professional development opportunities, and compensation. They insist that, in order to enhance teacher quality, it is important to develop fair and accurate teacher evaluation measures of teacher "effectiveness" and closely connect learning outcomes with accountability measures like salary, tenure, or dismissal.

Theoretical underpinnings. Scholars in this arena have suggested that teacher effectiveness should be measured by desirable outcomes that teachers are expected to produce, which was the progress of student achievement (Goldhaber, Harris, Loeb, McCaffrey, & Raudenbush, 2015; McCaffrey, Lockwood, Koretz, & Hamilton, 2003). This line of research assumed that teacher practice was comparable to a complicated black box, the essence of which was not easy for evaluators or school principals to understand. The principal-agent theory was one of the concepts used widely in this arena to understand an employer's performance (Heinrich & Marschke, 2010; Moynihan, 2008; Sun, Mutcheson, & Kim, 2015). This theory asserted that principals (employers) and individual agents (employees) had different levels of understanding about real performance. While agents knew their efforts and performance genuinely, principals inherently had a limited understanding of agents' true performance. Therefore, the theory proposed strategies of "coalignment of incentives," a system that aligned employees' performance and monetary reward. By doing so, managers could resolve organizational

issues and improve performance quality effectively without fully understanding true performance (Eisenhardt, 1989). In applying this "price mechanism of economics," schools would be able to manage teacher performance and quality effectively as they could align teacher evaluation criteria with student academic performance (Sun et al., 2015). From this perspective, teacher effectiveness should be measured by the degree to which teachers contribute to student learning, which is closely related to long-term economic values in a global knowledge society (Chetty, Friedman, & Rockoff, 2011; Goldhaber & Anthony, 2007; Hanushek & Rivkin, 2006; Rockoff & Speroni, 2011).

Value-added model and teacher accountability. To distinguish high from low performing teachers, scholars in this arena employed "…a collection of complex statistical techniques that use multiple years of students' test score data to estimate the effects of individual schools or teachers" (McCaffrey et al., 2003, p. xi). They named it the Value-Added Model (VAM) method of teacher evaluation, which has received enormous attention from policymakers and researchers because it provided theoretical and empirical grounds to include accountability based on student performance in teacher evaluations. Proponents of VAMs have argued that they enhanced the fairness and efficiency of the teacher evaluation system (Hanushek, 2003; Hanushek & Rivkin, 2006; Pennington & Mead, 2016). Compared to the traditional observational measures, VAMs were relatively objective in their ability to differentiate teacher quality and thus a useful tool to hold teachers accountable for their contributions to student learning. Further, they were reasonably efficient and feasible compared to other performance assessments (Pennington & Mead, 2016).

However, there have been growing concerns in recent years about national teacher evaluation reforms, including many teacher effectiveness measures like VAMs. Many scholars and practitioners have asserted that excessive reliance on VAMs made it difficult to measure the more comprehensive dimensions of genuine teacher performance (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012). Furthermore, many studies have found that, in reality, the federal-led teacher evaluation reform designed to increase teacher accountability has not led to improved student learning or educational equity (Alexander, Jang, & Kankane, 2017; Rothstein, 2010). For instance, Alexander and her colleagues analyzed state policies that included student achievement measures in teacher evaluations. They found that these accountability focused policies did not reduce the gap between white students and students of color across different states.

Effective teacher evaluation: multiple measures and feedback. Acknowledging the limitations discussed above, economists and human resource scholars have broadened their research scope in recent studies and recommended using multiple measures while retaining teacher effectiveness as the primary measure. They argued that using multiple measures could increase the reliability and validity of evaluation instruments, as different measures could evaluate different aspects of teacher performance (Harris, 2012). More importantly, these scholars have also turned their attention to professional development. Specifically, they have investigated how instructional information from teacher evaluations was used for professional development in addition to managerial tasks such as hiring, firing, promoting, and compensating teachers (Sun et al., 2015; Taylor & Tyler, 2012b). Their findings highlighted an

effective feedback process as a key mechanism to increase teacher effectiveness. The Gates Foundation's Measures of Effective Teaching (MET) project indicated that utilizing teacher effectiveness measures in combination with classroom observations and student surveys was effective not only in identifying effective teachers, but also providing valid and reliable feedback for teacher development (Kane, McCaffrey, Miller, & Staiger, 2013).

Teacher Education Theory

Unlike economists who have dealt with teaching quality issues from a macro perspective, scholars working in the teacher education arena have focused on teaching practices in classrooms, inside the "black box" of teaching and learning. These scholars have investigated how teachers learn and what elements constitute good teaching practice. On a theoretical level, their perspective stems from a social-cognitive (and sociocultural) approach that provides a comprehensive framework for examining the adult and social learning aspects related to teacher growth. From this perspective, teacher evaluation should be used as a valuable tool that can help teachers grow and improve their skills, knowledge, and practices as professionals (Darling-Hammond & Bransford, 2005). Researchers working in this arena define good teachers as professionals who have comprehensive knowledge, excellent teaching skills, and professional commitment (disposition) to help every child succeed (Shulman, 1998). As such, their definition of quality teachers is more comprehensive and holistic compared to that of economists.

Theoretical underpinnings. Scholars in this discipline have advocated for the overhaul of conventional teacher evaluation systems. Their concerns have mostly related

to a lack of consistency across local agencies and standards that guide teaching practices. They have criticized the fact that teacher evaluation has been conducted mainly based on easy-to-observe practices, such as classroom management, while, in practice, the discussion of ways to improve teaching practices and student outcomes was omitted. Limited resources and support for principals were also problematic: principals, who served as evaluators, received insufficient training to be competent evaluators and had little time to evaluate teachers' performance thoroughly.

Scholars in this realm have acknowledged the complex nature of good teaching. Teaching practices involve various aspects, including teachers' "moral and ethical ideal," theoretical grounds, skills, strategies, "judgment under uncertainty," "learning from experiences," and support from a professional community (Shulman, 1998, pp. 516-520). Accordingly, there can be no single right way to be a successful teacher. Some leading scholars in teacher-education theory have attempted to advance this notion, focusing on research related to common practices to improve student learning within a wide range of variations (Darling-Hammond & Bransford, 2005; Knight et al., 2015). That is, after conducting and reviewing extensive studies on the practices of competent veteran teachers, scholars identified key strategies that were commonly shared by experienced educators and subsequently developed a set of standard performance expectations for quality teaching.

Professional standards and performance assessment. Teacher education scholars have, therefore, suggested that teachers be evaluated according to professional standards using a comprehensive and wide range of evaluation instruments. They have

argued that this type of assessment could provide both formative and summative information on teacher performance. The most widely used standards-based teacher performance evaluations have shared important characteristics (Caughlan & Jiang, 2014; Danielson, 2008, 2013; Darling-Hammond, 2013; National Board for Professional Teaching Standards, 2002; Sato, 2014; Sato, Wei, & Darling-Hammond, 2008). Standards-based performance assessments encourage teachers to set their own goals and collect evidence of student learning. This way of teacher evaluation has enabled schools to establish a subject-specific, performance-based evaluation system that is not bounded by the limits of state-standardized tests, which only provide information about student performance in core subjects. Furthermore, scholars who advocated for standards-based performance assessments highlighted that this type of teacher evaluation adopts multiple measures through which teachers can receive rich feedback to guide classroom teaching practices. As a result, teachers and evaluators have more opportunities to participate in a comprehensive and authentic evaluation process compared to traditional paper-and-pencil tests or checklist forms of on-the-job evaluation

Effective teacher evaluation policy: Interactions and feedback. Recent research in the teacher education realm has acknowledged that a well-designed evaluation instrument alone is insufficient to result in teacher professional growth if school contexts are not considered. Scholars have proposed that the evaluation system should be implemented via collegial relationships among teachers that enable productive instructional conversations during the evaluation process in order to improve teacher learning (Darling-Hammond, 2013). Lee (2016) further argued that “By interacting with

other people, with the cognitive and physical artifacts available through human history, and with a focus on people's participation with practices within and across spaces, people engage in problem-solving and acts of learning" (p. 76). This highlights the importance of the social aspects related to what and how information is shared among teachers.

Organizational Theory

Organizational theorists have also provided valuable insights into teacher evaluation. In contrast to the two areas of scholarship described above, researchers in this arena are less interested in directly addressing what constitutes valid and reliable teacher evaluation measures. Rather, they focus on the implementation of evaluation policies which include understanding the process of teacher evaluation, its impact on teachers and students, and the role of local policy actors in shaping implementation. This approach provides two meaningful perspectives on teaching quality and teacher evaluation policy. First, it views teaching quality as an organizational component that interacts with multiple factors, rather than individual teacher traits. Second, it draws attention to the policy implementation *process* of teacher evaluation and its impact on teachers and teaching.

Theoretical underpinnings. Scholars who take an organizational approach have criticized traditional approaches to teaching quality and teacher evaluation that they view as overly focused on individuals. They have argued that teacher quality has been operationalized in the existing literature based on three categories of individual teacher characteristics: (1) cognitive resources (i.e., knowledge, beliefs, attitudes, and dispositions); (2) classroom performance; and (3) contributions to student learning

(Kennedy, 2008, as cited in Knight et al, 2015). Kennedy (2010) observed that this is an "attribution error" of traditional narratives of teacher quality, suggesting that "Teacher effects are not as stable from year to year as we would expect them to be if they were due primarily to enduring qualities within teachers themselves" (p. 592). Synthesizing numerous studies, she showed that student learning was influenced not only by teacher characteristics but also situational contexts, including the amount of planning time, teaching materials, working conditions, organizational climate, institutional practices, and student characteristics (e.g., their moods, prior knowledge, diverse background, etc.).

From this viewpoint, the concept of teaching quality encompasses the situational, organizational, cultural, and collective nature of teaching (Cohen, 2010; Hiebert & Morris, 2012; Kennedy, 2010). This line of research counters the conventional assumption that improving teachers' skills guarantees the improved quality of teaching and student learning. Instead, it accounts for how the interaction between teachers and contexts influences student learning and how educational systems can support quality teaching.

Implementation of teacher evaluation policy. In this arena, scholars have investigated the interactions among teacher evaluation measures, policy, and organizational dynamics to broaden our understanding of the teacher evaluation implementation process. They have argued that, to implement policies as intended, local actors must understand the purpose of policies and develop the capacity to implement them. However, local actors often misunderstand purposes or cannot implement them successfully. For example, Ballou and Springer (2015) found that several teacher

evaluation systems that used VAMs as primary evaluation criteria ignored potential estimation errors included in the complex modeling of VAMs. This misled school leaders when they considered personnel decisions like hiring, dismissal, or retention. As a result, teachers were deemed ineffective unfairly based on factors other than student learning. From this perspective, teacher evaluations that are gauged using complex statistical models such as VAMs are often misunderstood and, thus, yield serious evaluation errors.

Scholars have also found that school adminisraters' and educators' trust and belief in measures were another critical factor that influenced teacher evaluation implementation in practice (Harris & Herrington, 2015; Jiang, Sporte, & Luppescu, 2015). Studies showed that teachers and principals believed that observational instruments better reflected authentic teaching practices and were more relevant and informative than VAMs. Teachers also wanted to be able to identify their strengths and weaknesses more precisely in order to inform their own improvement. Both leaders and educators found that observational instruments provided the types of information that they sought, but not VAMs. As Harris & Herrington (2015) suggested, the distrust of VAMs had serious implications for how teacher evaluation results were interpreted and used.

Effective teacher evaluation: Professional community and feedback. Scholars who take an organizational perspective argue that teachers' work and professional learning are shaped by their school's collective assumptions, norms, and practices. Research had shown that teaching practices were more likely to be improved when teachers had opportunities to reflect on their instruction through active interactions and

conversations with their peers and school leaders. For instance, researchers conducted a large set of qualitative studies with more than 800 teachers in multiple states to understand principals' everyday practices that enhanced classroom instruction (Blase & Blase, 1999). Two major practices emerged: (1) talking with teachers to promote reflection and (2) promoting professional growth. Strategies associated with the first theme of talking with teachers to promote reflection include the "…principal strategies of making suggestions, giving feedback, modeling, using inquiry and soliciting advice and opinions from teachers, and giving praise" (p. 367).

The feedback that teachers receive from their peers is also essential for professional learning. Many scholars researched the role of professional learning communities in supporting teaching practices and school-wide instructional improvement. Stoll and Louis (2007) defined the professional learning community as a group of teachers who share their practice critically in "…an ongoing reflective, collaborative, inclusive, learning-oriented, growth-promoting way" (p. 2). Professional communities create social networks for teachers by engaging them in intellectual conversations and allowing them to share their expertise in classroom practice. These networks also enable teachers to leverage social capital, which is defined as the resources that an individual can draw on through social relationships, to gain valued outcomes (Coburn & Russell, 2008; Mulford, 2007; Spillane, Hopkins, Sweet, & Shirrell, 2017).

Based on this long-standing research on the school environment and social aspects of teaching that shape and influence teacher learning, organizational theorists argue that successful teacher evaluation depends on school climate and interactions

among key policy actors. The later sections of this chapter provide more details regarding what organizational factors should be accounted for in order to grow professional capacity through teacher evaluation.

Summary of Theoretical Perspectives

In the previous sections, I reviewed the theories and epistemologies behind the discussions on teacher evaluation policy and research. Figure 2.1 summarizes the three major scholarly approaches to teacher evaluation policy. Firstly, scholars in economics and human resources theory examine teacher evaluation policy at the macro-level, focusing on the societal and economic values of education and policy that holds teachers accountable for student outcomes. Their research has largely centered on the accuracy of teacher evaluation measures that gauge teachers' contributions to student learning, which is related closely to long-term economic values in a global knowledge society. They believe that teachers' instructional quality can be enhanced by holding teachers accountable for outcomes, using high-stakes measures such as dismissal, financial bonuses, tenure, etc.

In contrast to this first group of researchers, teacher education scholars investigate teaching practices and performance at the micro-level. Unlike economists, who view classroom teaching as a complicated black box, they attempt to define exemplary teaching and professional standards. From this perspective, professional growth is a key mechanism for enhancing teaching quality. As such, ideal teacher evaluation should provide information and feedback on teaching performance so that teachers can learn and grow professionally. Accordingly, they work to develop performance assessments based

on professional standards with multiple sources that can be used to provide ample feedback on teaching practices and opportunities to remedy weaknesses.

Unlike economists or teacher education theorists, organizational theory scholars are interested in the impacts of organizational dynamics and the teacher evaluation implementation process on teaching and learning in classrooms. From this perspective, both teaching quality and teacher evaluation policy are interconnected and influenced by various organizational contexts. While developing accurate and reliable teacher evaluation measures is important, teaching quality and teacher evaluation policy should be considered and discussed alongside these organizational contexts. Many studies highlight that the feedback that teachers receive from peers and school leaders through their interactions during the evaluation process contributes directly to teaching quality.

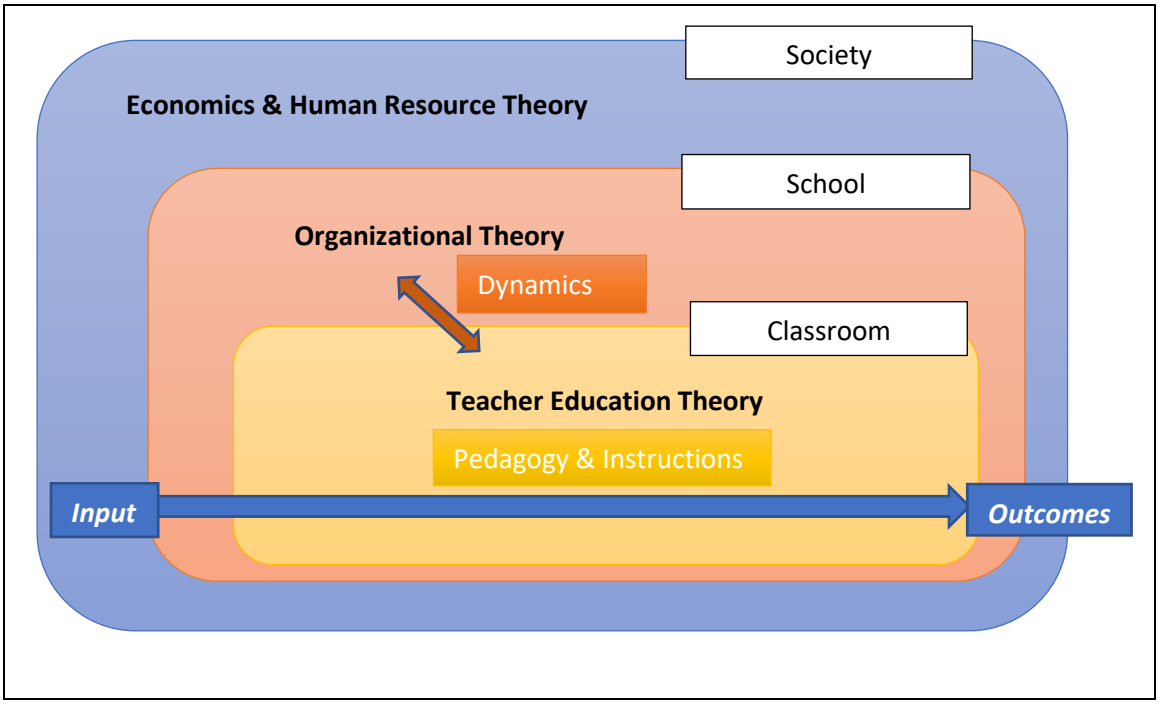


Figure 2.1 ***Different Theoretical Approaches to Teacher Evaluation Policy***

Despite their differences, recent studies using each of these approaches have had at least one consistent finding: teacher evaluation improves teaching and learning when it provides ample professional growth opportunities. Researchers across disciplines have shown that a rigorous teacher evaluation system that uses detailed feedback to provide rich information about *how* to enhance teaching can help both novice and experienced teachers improve their teaching practices (Darling-Hammond, 2013; Hallgren et al., 2014; Smylie, 2014; Taylor & Tyler, 2012b). These recent findings have steered policy and research on teacher evaluation away from mere accountability and toward professional growth. Leading scholars across different disciplines have reconceptualized the purpose of teacher evaluation to encompass the dual goals of teacher accountability and professional growth and have highlighted the importance of school climate to achieve these dual goals.

Synthesis of Recent Literature

Dual Purposes of Teacher Evaluation Policy

Synthesizing numerous and comprehensive studies, leading scholars argue that teacher evaluation should achieve both the goals of accountability and improvement/professional learning (Darling-Hammond, 2013; Hargreaves & Braun, 2013; Murphy et al., 2013; Papay, 2012; Plecki, Elfers, & Yeh, 2015; Smylie, 2014; Taylor & Tyler, 2012b). Based on a review of the literature, Plecki et al. (2015) defined the difference between accountability and improvement evaluation systems as follows:

> Accountability systems have primarily focused on using teacher evaluation to make decisions about hiring, firing, tenure or salary. In recent years, evaluation

> for accountability purposes has included measures of how teachers contribute to student learning. This implies a high-stakes system of evaluation ... In contrast, evaluation for improvement uses the process to inform decisions about the kinds of professional learning opportunities needed to help teachers and schools engage in continuous improvement. (p. 3-4)

Notably, recent research showed that a major factor weakening the effects of teacher evaluation reforms was the separation between teacher evaluation and professional development (Papay, 2012; Smylie, 2014). Smylie (2014) criticized the fact that most teacher evaluation systems "…give professional development short shrift, make vague and weak provisions for professional development, or leave it to individual teachers or their schools and school districts to make such linkages themselves" (p. 98). For teacher evaluations to be truly effective, their results should be reflected in teacher professional development.

The concerns about and unintended consequences of a high-stakes, accountability-oriented teacher evaluation system have been well-documented. First, studies have shown that such teacher evaluation did not achieve its intended goals of improving student learning (Alexander et al., 2017; Rothstein, 2010). Also, many studies have shown unintended consequences of high-stakes policies, both within school organizations and in complex labor markets, including narrowing curricula, gaming the system (cheating), and high teacher turnover rates in disadvantaged schools, to name a few (Goldhaber, 2015; Hargreaves & Braun, 2013; Harris & Herrington, 2015).

Accordingly, recent research on teacher evaluation policy has reconceptualized the purpose of teacher evaluation, highlighting both accountability and professional growth. As seen in the policy analysis of countries like Finland, South Korea, and Japan, efforts to improve teacher accountability do not necessarily result in linking teacher evaluation outcomes with high-stakes consequences such as dismissal or non-renewal of contracts. A growing number of scholars argued that teacher evaluation should provide meaningful and authentic guidelines for classroom teaching to create ample opportunities for professional learning opportunities. Numerous studies about teacher evaluation across three disciplines emphasize that a rigorous teacher evaluation system that uses detailed feedback to provide rich information about how to enhance teaching could help both novice and experienced teachers improve their teaching practices (Darling-Hammond, 2013; Hallgren et al., 2014; Hanushek, 2003; Taylor & Tyler, 2012). Effective teacher evaluation provides constant feedback on teaching based on clear standards and opportunities for teachers to grow. Through a rigorous feedback loop, teachers and evaluators can participate in a more comprehensive and authentic assessment process compared to traditional paper-and-pencil tests or checklist forms of on-the-job evaluation. Taylor and Tylor (2012) suggested that teacher evaluation should be used as an "information mechanism" using "performance appraisal as an integral part of long-run employee development rather than as a tool in a rewards-and-punishment incentive scheme" (pp. 3629 - 3630). These features also aligned with the beliefs of school leaders and teachers on the role of evaluation (Goldring et al., 2015) and acknowledged the professional achievements of high-performing teachers.

Shared Responsibility for Effective Implementation of Teacher Evaluation Policies

Effective teacher evaluation is intended to result in quality teaching and learning. For teacher evaluation to lead this outcome, teacher evaluation policy should achieve both the goals of teacher accountability and professional growth, as outlined above. Nevertheless, the attempt to achieve both purposes often faces enormous challenges. Hargreaves & Braun (2015), studying various policy reforms in several sectors, noted that there were contentious tensions and even direct conflicts between these two goals. They elaborated that the conflicts between improvement and accountability "are most likely to be resolved when there is collaborative involvement in data collection and analysis, collective responsibility for improvement, and a consensus that the indicators and metrics involved in data-driven improvement accountability are accurate, meaningful, fair, broad and balanced" (p. i).

Darling-Hammond (2013) proposed a similar solution to this issue. She also argued that collective responsibility and a collaborative decision-making process were necessary to produce better personnel decisions than traditional teacher evaluation while simultaneously providing professional learning opportunities. Peer Assistance and Review (PAR), which has been implemented in various states and districts, is a good example. First, PAR provides extensive learning opportunities for teachers to improve instruction through collaborative consulting with teacher mentors and peers based on common teacher evaluation frameworks. Regarding teacher accountability, the PAR governing body, teachers' unions, and school board members made key personnel decisions together while going through teacher evaluation results. When teachers did not

meet the evaluation criteria, they proposed a supporting plan to remedy their weaknesses. If they still did not meet the professional standards after receiving assistance, the PAR governing board dismissed those teachers. This collaborative decision-making process created a "shared sense of responsibility" and a "focus on teaching and learning" among union officials and district administrators (p. 123). Under PAR, teacher performance, especially those who received the intervention, improved considerably, and both retention rates and trust in evaluation decisions increased.

In sum, recent research highlights that teacher evaluation can be an empowerment tool when implemented in a collaborative school climate and shared responsibility among local policy actors. This finding counters the traditional perspective on teacher evaluation, namely that an individual teacher is responsible for their teaching and must be held accountable for the outcomes. These studies show that teaching performance is interconnected with organizational contexts and, thus, teacher evaluation should be implemented with a shared responsibility among the school community to leverage teacher capacity. Indeed, teacher evaluation is a social and organizational endeavor, which calls for examining teacher evaluation policy in organizational contexts.

Effects of Teacher Evaluation on Professional Capacity

Other research findings have indicated that teacher evaluation, when implemented as outlined above, can produce positive impacts on various aspects of teachers and teaching. These effects are not limited to improving teaching practices but also include increasing teacher motivation, confidence, and job satisfaction and promoting teacher leadership. Research evidence has shown that evaluation geared towards both

accountability and professional development and implemented under conditions of shared responsibility accelerated teacher learning in pedagogical knowledge and skills. When teachers received critical and detailed feedback through such a rigorous teacher evaluation system, their teaching skills and practices improved considerably (Hallgren et al., 2014; Sato et al., 2008; Sun et al., 2015). For instance, Taylor and Tyler (2012) found that teachers' performance increased under a teacher evaluation model "whereby teachers learn new information about their own performance during the evaluation and subsequently develop new skills, or increase long-run effort, or both" (p. 3629). Mid-career teachers, who were typically unlikely to change their practices after several years in the classroom, also improved their effectiveness when undergoing a rigorous teacher evaluation process that provided a detailed and multi-faceted review. Teachers who participated in teacher evaluation were even more productive in later years after evaluation.

Moreover, the effects of professional learning opportunities and collaborative networks on teacher motivation, confidence, and job satisfaction are well-documented (Ames, 1990; Canrinus, Helms-Lorenz, Beijaard, Buitink, & Hofman, 2012; Davis & Wilson, 2000; Durksen, Klassen, & Daniels, 2017; Thoonen, Sleegers, Oort, Peetsma, & Geijsel, 2011). Findings show that these positive emotional effects are also relevant for teacher evaluation when it is coupled with improvement opportunities in a collaborative and supportive climate. Firestone (2014) examined motivation theories to understand how teachers' intrinsic and extrinsic motivation was developed depending on different evaluation approaches. He concluded that teacher evaluation should provide teachers

with useful feedback and learning opportunities to maintain their motivation for effective teaching. Beerens (2000) also provided empirical evidence that teacher evaluation motivated teachers for continuous learning when the evaluation system supported professional learning by incorporating constructivist teaching, education research findings, and reflective practice, and by promoting collaboration through peer coaching and evaluation.

In addition, how teacher evaluation is implemented significantly affects teacher job satisfaction and stress. While high-stakes, accountability-oriented evaluation decreased job satisfaction, a supportive and collaborative evaluation process may increase job satisfaction and improve retention (Darling-Hammond, 2013). Ford, Urick, and Wilson (2018), who examined 2013 TALIS data, found a small, but significant relationship between the perception of supportive teacher evaluation experiences and U.S. secondary teachers' satisfaction. Teachers who felt that their evaluation resulted in positive changes in their teaching were also more likely to be satisfied with their jobs. Moreover, motivation, job satisfaction, and confidence are important factors that affect teacher performance and teacher's experience, and, at the same time, these are empirically indistinct in the empirical literature (Canrinus et al., 2012; Davis & Wilson, 2000; Judge & Bono, 2001; Pearson & Moomaw, 2005).

Research has also provided evidence that a supportive and participatory teacher evaluation process led to teacher leadership. A school climate of shared responsibility and collaboration created more opportunities for teachers to grow their professional capacity by allowing them to participate in school initiatives and decision-making

processes. Shared responsibility and collaboration also enabled them to influence their colleagues to improve educational practices (Harris, 2003; Printy, Marks, & Bowers, 2010; York-Barr & Duke, 2004). In this sense, Goldstein (2010)'s study on PAR sheds light on how teacher evaluation could be associated with teacher leadership. Findings showed that teachers showed a substantially higher level of accountability with PAR than before the program was implemented. As a result of the program, they took ownership of their work and felt that they could influence others.

Research Gaps

As has been laid out above, linking teacher evaluation with professional growth outcomes can increase the scope and degree of its positive effects. Thus, teacher evaluation should be understood as a professional growth tool that induces more opportunities for improving professional capacity more widely. In return, expanding professional capacity has long-term effects on teaching practices, teacher retention, and school improvements.

When positing teacher evaluation as a mechanism for professional growth, the organizational contexts that will support such mechanisms must also be considered. However, although teacher evaluation has been at the center of the policy and research discussion globally for decades, organizational factors related to teacher evaluation have not been studied in a systemic way. More studies using a comprehensive, organizational approach to examine the educative purpose of teacher evaluation (i.e., providing ample feedback and opportunities for professional growth) in organizational contexts are needed. In particular, a dearth of studies investigated the association among teacher

evaluation policy implementation, school climate, and their relationships with teacher professional capacity using quantitative methods. This section summarizes the specific research gaps in the current teacher evaluation research in four main areas, which ground the research questions of this paper.

First and foremost, while policies have centered on teacher evaluation as a policy lever to uphold teacher responsibility for student achievement, less consideration has been given to the educative function of teacher evaluation. Since the early 1990s, the U.S. federal government has chosen to prioritize accountability over professional development by increasing regulatory standardized monitoring (Cohen-Vogel, 2005; Plecki & Loeb, 2004). Many researchers, thus, focused on how education policy could hold teachers accountable for student learning through reliable and efficient teacher evaluation measures. However, as established above, a growing body of research shows that teacher evaluation needs to be thoughtfully designed and implemented to provide professional growth opportunities along with the purpose of teacher accountability, so that it can be an effective tool to enhance classroom practice (Darling-Hammond, 2013; Murphy et al., 2013; Smylie, 2014). Teacher evaluation needs to provide ample feedback and opportunities for reflection that stimulates self-directed improvement, leading to improving the center of teaching and learning. Furthermore, numerous studies have shown that teacher evaluation implemented solely focused on teacher accountability, without considering professional growth, is not effective at enhancing teaching and learning or reducing learning gaps. It even can have unintended consequences both

within school organizations and in complex labor markets. Thus, more research is needed on the professional growth aspect of teacher evaluation.

Second, the literature on teacher evaluation, while robust, has seldom considered organizational contexts that promote and support teacher professional growth through evaluation policy implementation using quantitative methods. Harris and Harrington (2015) keenly pointed out that, since the inception of Obama's Race to the Top in 2009, research has focused almost entirely on the statistical properties of the validity and reliability of evaluation measures. However, such studies tell us "very little about the effects on teaching and learning that come from embedding value-added into policies like teacher evaluation, tenure, and compensations" (p. 71). Highlighting the importance of understanding the policy implementation process, Harris and Harrington (2015) further argued that educators' perceptions shape their behavioral responses to teacher evaluation, which, in return, is closely related to its effects on teaching and learning. However, research remains scant on how policy should be implemented to enhance teaching and learning, and, further, how school climate or organizational elements could support the successful implementation of teacher evaluation policy. Research examining teacher evaluation and its impact in an organizational context is growing, but thus far, empirical studies have largely taken a qualitative approach. While these findings are valuable, much information is still needed to understand how these variables are related to professional capacity.

Third, the effects of teacher evaluation remain narrowly investigated in the literature. This seems to be mainly because the policy and research narratives on teacher

evaluation have been dominated by the one-dimensional approach of evaluation's impact on student achievement. When the purpose of teacher evaluation is limited to accountability for student achievement, the effectiveness of evaluation can only be assessed by measuring student learning gains. Even though recent findings using all three scholarly approaches to the subject imply more comprehensive effects of teacher evaluation on professional capacity, few studies have comprehensively examined different facets of teaching beyond student learning outcomes, such as daily classroom practice and the emotional and social aspects of teaching.

Lastly, little attention has been paid to the global narratives behind the development of educational policy. Although teacher evaluation at the local level is often considered to be a local issue, Akiba (2017) argued that global discourses on teacher quality influenced the development and implementation of teacher evaluation and compensation reforms within several countries. In this regard, he asserted that "few studies systematically analyzed what explains the cross-national difference in how a national, federal, or state government develops and implements a teacher reform influenced by global dynamics" (p. 153). Given the significant influence of global dynamics that have promoted the importance of holding teachers accountable, it is critical to investigate how nations have implemented teacher evaluation policies similarly or differently in response to this global trend, as well as teachers' perceptions of their effectiveness.

Research Questions

The purpose of this study is to provide a new perspective on effective teacher evaluation and how it is implemented by policymakers and school leaders. By emphasizing both accountability and professional development as dual goals of teacher evaluation, this study addresses the need to attend to how it is implemented and used on a local level. Particularly, this study examines the implementation of teacher evaluation and feedback systems at the school level in four high-performing countries: the United States, Finland, Korea, and Japan. The study also accounts for the research findings that teachers' work and their professional learning are shaped by the organizational environment in which their schools are situated (Lortie, 1975; Waller, 1932). Organizational decision-making and learning processes are complicated and influenced by elements within and outside of organizations (Bolman & Deal, 2013; Morgan, Gregory, & Roach, 1997; Shafritz, Ott, & Jang, 2011; Sutton & Levinson, 2001). As a result, these fissures of educational organizations require a multidimensional approach in order to understand teachers' work in depth. Theoretically oriented by organizational research, this study investigates the following research questions:

(1) How are national teacher evaluation policies implemented at the local level in four countries?

   a. To what extent are teachers formally evaluated? Who evaluates teachers, and what evaluation methods are used?

b. How do principals and teachers perceive the outcomes and impact of teacher evaluation and feedback?

(2) How are teacher evaluation policies and school climate associated with teachers' perceived professional capacity in four countries?

a. To what extent is teacher evaluation associated with professional capacity when its outcomes support teacher professional growth?

b. To what extent is the school climate of shared responsibility associated with professional capacity?

(3) How are teacher evaluations associated with teachers' professional capacities when evaluation is accompanied by support of teacher professional growth?

Using well-established international survey data, the OECD TALIS 2013, allows for an examination of this question in both the U.S. and other countries. This research topic is timely, as ESSA recently encouraged more flexibility for teacher evaluation (National Education Association, 2015; Sawchuk, 2016), therefore providing opportunities to rethink effective teacher evaluation policy.

Significance of the Study

On a theoretical level, examining how teacher evaluation is implemented will provide insight into effective strategies for leaders who hope to improve teacher evaluation in their schools. This study focuses on two key factors of successful teacher evaluation policy implementation: teacher evaluation outcomes linked to professional

growth and shared responsibility in schools. Further, current studies of teacher evaluation policy have narrowly viewed the scope of impact of teacher evaluation. This paper takes one step further and investigates the relationships between teacher evaluation and professional capacity by looking at its role more broadly.

Methodologically, this study explores the relationships between teacher evaluation outcomes and professional capacity using a quantitative approach. Researchers have made efforts to understand how teacher evaluation increases teacher performance and student achievement (see Taylor & Tyler, 2012). While these studies were robust, they took an individualistic approach to teaching, missing a critical dimension: the organizational aspects of teaching. This paper advances and expands this model by adding factors related to teacher evaluation policy and school climate.

Finally, few have conducted a comparative study of how teacher evaluation policies have been implemented across several nations. Analyzing cross-national similarities and differences in teacher evaluation policy and its implementation can deepen our understanding of how such approaches to teacher evaluation might affect other initiatives to improve teacher quality.

Chapter 3 Methods

Data Source and Collection

To understand the relationship between key factors affecting teacher evaluation and the enactment of teaching across national education systems, this study uses the TALIS 2013 teacher and principal survey data. The "core" target population for TALIS 2013 was teachers and school leaders in lower secondary education in 34 nations (OECD, 2014). TALIS 2013 adopted a two-stage stratified sampling design[2]. This means that schools were randomly selected within countries, and teachers were randomly selected within those schools. Two questionnaires, one for school leaders and another for teachers, were completed in each school. All participating couturiers were mandated to survey the population of lower secondary level, while each could choose to survey at the primary and upper secondary levels. The survey also targeted the general schools and their educators, exclusive of schools that solely serve students with disabilities; substitute or emergency teachers; teachers on long-term leave; and teachers of adult education courses.

I have selected four countries for comparative analysis: the U.S., Japan, Korea, and Finland. The non-US countries selected for this study are consistently ranked above the international average in reading, math, and science PISA, and are known for their stable teacher hiring and support systems. The 2013 data provides a fertile ground to examine teacher evaluation policy and its impact. The global policy debates centered

[2] According to the OECD TALIS technical report (2013), "stratification resulted in a combination of some or all of the details relating to geography, source of financing, type of educational program and school size" in most cases (p.78). The TALIS 2013 U.S. technical report (Strizek, Tourkin, & Erberber, 2014) provides further information about the U.S. data. The teacher-level survey used the stratification variables of year of birth, gender and main subject domain. For the school level survey, school type (public/private), grade level, urbanicity, region, and percent minority students were used as stratification variables.

around teacher quality began in the 2000s, heightened in the 2010s, causing policymakers in all four of these countries to actively implement various teacher evaluation reform initiatives between 2000 and the early 2010s, and thus making 2013 data appropriate for gauging the aftermath of these initiatives. Understanding the implementation of teacher evaluation policies in these three nations will provide valuable insights for educators and policymakers in the U.S., where efforts to experiment with teacher evaluation have emerged more recently.

Sample Sizes, Participation Rates, and Caveats

This study uses samples of schools and teachers designed to be representative of each country's lower secondary school teachers (OECD, 2014; Strizek et al., 2014). The sample sizes and participation rates in the TALIS 2013 survey of these four nations are summarized in Table 3.1.

Table 3.1 ***Participation and Estimated Size of Teacher Population, ISCED Level 2, 2013***

	Number of participating schools	Responding teachers in participating schools	School participation before replacement (%)	School participation after replacement (%)	Teacher participation in participating schools %)	Overall participation (%)	Weighted estimated size of teacher population
Finland	146	2, 739	90.5	98.6	91.3	90.1	18, 386
Japan	192	3,484	88.0	96.0	99.2	95.2	222,809
Korea	177	2,933	68.3	88.9	88.1	78.3	85, 184
U.S.	122	1,926	39.4	61.6	83.3	51.4	1, 052, 144

Source: 2013 TALIS technical report.

The TALIS Board of Participating Countries set data standards to ensure valid and reliable comparisons across different countries. These standards required that each country's dataset have valid responses from at least 50% of original schools surveyed and at least 75% of all sampled schools (after replacement). In addition, at least 75% of teachers within a selected school should participate in the survey to ensure a minimum

level of reliability. Table 3.1 shows that Finland, Japan, and Korea all meet this international standard for a comparison study.

Although the U.S. participation rate has not met the TALIS technical standards required for international comparison, the report also indicated that this rate was deemed sufficiently high to report the U.S. data independently (OECD, 2014). According to the U.S. TALIS technical report, "the TALIS Board agreed that the U.S. response rate and quality of collected data were nonetheless of sufficiently high quality to report based, in part, on an initial nonresponse bias analysis conducted by the United States and submitted to the OECD for consideration" (Strizek et al., 2014, p. 33). Thus, although the U.S. did not meet the response rate requirement, the quality analysis indicates that the inclusion of the U.S. in this study was warranted.

The purpose of this study is to understand how teacher evaluation policy is associated with impacts of teacher evaluation as well as school climate within each country, rather than comparing the survey results directly. This study does not aim to rank each system based on survey data but to analyze the patterns of the dynamics of key factors pertinent to teacher evaluation and feedback in each system. However, the U.S. data are shown separately from the other participating education systems that achieved acceptable response rates following the guidelines from the U.S. TALIS technical report.

Sample Description: Gender, Age, Years of Experience of Participants

Table 3.2 shows the percentage of lower secondary education teachers in the sample by gender, average age, and average years of experience. A majority of teachers were female in Finland, Korea, and the U.S., while in Japan the percentage of male

teachers was higher than female teachers. The average ages of teachers were similar across all countries, ranging from 42 to 44. The average years of teaching experience ranged from 14 to 17 years.

Table 3.2 ***Gender, Average Age, and Average Years of Working Experience as a Teacher in Total: Lower Secondary School Teachers, 2013***

	Female		Average age		Average years of working experience as a teacher in total	
Education system	Percent	(S.E.)	Average	(S.E.)	Average	(S.E.)
Finland	72.4	(0.75)	44.1	(0.23)	15.5	(0.23)
Japan	39.0	(0.80)	41.9	(0.24)	17.4	(0.23)
Korea	68.2	(1.07)	42.4	(0.28)	16.4	(0.31)
United States	64.4	(1.06)	42.2	(0.39)	13.8	(0.41)

Source: the U.S. TALIS technical report

Analytical Strategies

Descriptive Analysis

To understand teacher evaluation policy at the local level in all four countries, which was also the first research question, I analyzed the descriptive statistics of various aspects of teacher evaluation policy using both principal and teacher evaluation surveys. Estimation for surveys with complex designs like TALIS requires special attention (Becker, Dumais, LaRoche, & Mirazchiyski, 2016; Heeringa, West, & Berglund, 2017; OECD, 2014). All data in the TALIS database were collected from random samples of schools and teachers, and the random samples accounted for not only the sampled schools and teachers, but the entire educational system. Given this sampling design, there were different selection probabilities for sampling schools and teachers within selected schools. In consideration of the disproportional selection probabilities among the schools and teachers, scholars suggest that the analysis of complex surveys should use sampling weights that reflect and compensate for the unequal selection probabilities. The

descriptive analysis was conducted based on the survey data which were adjusted by appropriate weights suggested in the 2013 TALIS user guide (Becker et al., 2016). These results were analyzed through SPSS software, which provided multiple features to conduct complex survey analysis. When the data of interest were available from the OECD TALIS results website (2014), I adapted the data tables and further analyzed them to answer the research question pursued in this study.

Data Reduction

To investigate the associations among key variables, it was critical to develop psychometrically valid measures of teacher evaluation, school climate, and the positive impact on teaching across all four countries. Teacher survey data were used to construct key variables. The data were first examined with exploratory factor analysis. Then, key latent variables for this cross-national study were identified through multi-group Confirmatory Factor Analysis (CFA) and a follow-up measurement invariance analysis. CFA helps develop scales that consist of multiple items representing certain characteristics, resulting in higher reliability and validity (OECD, 2014). These scales are advantageous for measuring conceptual characteristics like beliefs, attitudes, or practices, which are the variables of interest in this study. CFA can also alleviate multicollinearity issues among variables. The measurement invariance test was used as a follow-up to validate cross-country statistical comparisons and was conducted at three levels: configural, metric, and scalar.

The study includes three latent variables: SR, *Shared responsibility*; EFP, *Evaluation and feedback outcomes linked to professional growth*; PC, *Positive impact of*

feedback on professional capacity. The PC is a second-order latent variable that encompasses three subscales including impact on teaching practices, teacher motivation/job satisfaction/confidence, and teacher leadership. All these latent variables emerged and were validated through the multi-group CFA. The results of the multi-group CFA analysis are elaborated in the findings section. For the rigor of this study, only latent variables constructed with three or more variables and meeting the standards of the research were selected (Kline, 2015). The reliability measures of scaled items are included in the Appendix A.

Shared Responsibility. The SR variable was included in the analysis to link the implementation of teacher evaluation and school climate. This variable was comprised of five indicators that represent a collaborative school culture and the participation of multiple stakeholders. This construct was built using five items such as "This school has a culture of shared responsibility for school issues", "This school provides parents or guardians with opportunities to actively participate in school decisions", etc. (see Table 3.3). All items were measured on a four-point scale where 4 is "strongly agree". High values indicated a high likelihood that teacher evaluation and feedback outcomes supported teachers' professional growth occurred.

Table 3.3 ***Measured Items for Shared Responsibility***

Scale	Variable	Item Wording
Shared responsibility	TT2G44A	This school provides staff with opportunities to actively participate in school decisions
	TT2G44B	This school provides parents or guardians with opportunities to actively participate in school decisions
	TT2G44C	This school provides students with opportunities to actively participate in school decisions
	TT2G44D	This school has a culture of shared responsibility for school issues
	TT2G44E	There is a collaborative school culture which is characterised by mutual support

Teacher evaluation and feedback outcomes for professional growth. The EFP variable encompasses the degrees of teacher evaluation outcomes that occurred in each country, particularly outcomes related to teacher professional growth. It was comprised of four indicators representing teacher evaluation outcomes that led to teacher learning and support. This construct was built using four items such as "A development or training plan is established for teachers to improve their work as a teacher" (see Table 3.4). All items in the scales were measured on a four-point scale where 4 is "strongly agree". High values indicated a high level of the likelihood that teacher evaluation and feedback outcomes that support teacher professional growth occurred.

Table 3.4 ***Measured Items for Teacher Evaluation Outcomes for Professional Growth***

Scale	Variable	Item Wording
Teacher evaluation and feedback outcomes for professional growth	TT2G31D	A development or training plan is established for teachers to improve their work
	TT2G31E	Feedback is provided to teachers based on a thorough assessment of their teaching
	TT2G31G	Measures to remedy any weaknesses in teaching are discussed with the teacher
	TT2G31H	A mentor is appointed to help the teacher improve his/her teaching

Positive impact on a professional capacity. The PC variable was measured on three subscales including the perceived impact on teaching practices, teacher motivation/confidence/job satisfaction, and teacher leadership. The wording of the survey question was, "Concerning the feedback you have received at this school, to what extent has it directly led to a positive change in any of the following?" Measurement items were categorized into three latent constructs:

- Perceived impact on teaching practices: This latent construct was measured by five items, such as "Your teaching practices".
- Perceived impact on teacher motivation/confidence/job satisfaction: This included three items, such as "Your confidence as a teacher".
- Perceived impact on teacher leadership: This construct was built on five items, such as "Your role in school development initiatives".

See table 3.5 for detailed information about survey items. All items in the scales were measured on a four-point scale where 1 is "No positive change" and 4 is "A large change". High values indicated the high level of positive change resulted from the teacher evaluation and feedback system at the local level.

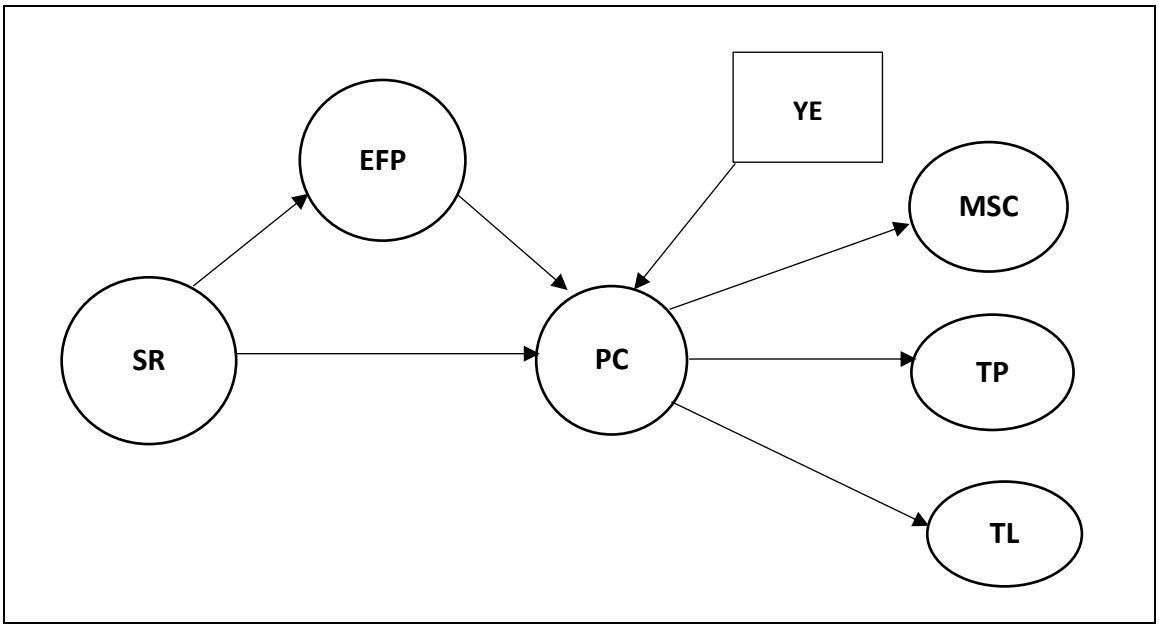
Table 3.5 ***Measured Items for Teacher Evaluation/Feedback Impact***

Scale	Subscale	Variable	Item Wording
Positive impact on professional capacity	Teaching practices	TT2G30H	Your classroom management practices
		TT2G30I	Your knowledge and understanding of your main subject field(s)
		TT2G30J	Your teaching practices
		TT2G30K	Your methods for teaching students with special needs
		TT2G30L	Your use of student assessments to improve student learning
	Confidence/ Job satisfaction/ motivation	TT2G30F	Your confidence as a teacher
		TT2G30M	Your job satisfaction
		TT2G30N	Your motivation
	Teacher leadership	T2G30A	Your public recognition from the principal and/or your colleagues
		TT2G30B	Your role in school development initiatives (e.g. curriculum development group, development of school objectives)
		TT2G30C	The likelihood of your career advancement (e.g. promotion)
		TT2G30D	The amount of professional development you undertake
		TT2G30E	Your job responsibilities at this school

Factor Analysis and Structural Equation Modeling

Once the variables were validated for cross-national analysis, I further explored the associations among latent variables to understand the conditions to facilitate teacher education to improve teaching practices by employing Structural Equational Modeling (SEM). SEM is useful for analyzing structural relationships among multiple factors, including the unobservable latent variable (Kline, 2015). SEM is particularly useful for this study as it provides a tool to analyze the pattern of organizational and policy conditions. I specified the SEM model based on an extensive literature review across three scholarly origins (Figure 3.1). The model explores the direct and indirect effects among the three key variables of the shared responsibility, teacher evaluation for

professional growth, and the positive impact of teacher evaluation on a professional capacity. In this model, I also make the assumption that, although the TALIS data are cross-sectional, teacher's assessments of shared responsibility reflect a relatively stable feature of a school culture, while teachers' responses about evaluation and feedback events are more likely to reflect recent experiences.



Note. EFP = Evaluation and Feedback Outcomes for Professional Growth; MSC = Positive Impact on Motivation, Job Satisfaction, and Confidence; PC = Professional Capacity; SR = Shared Responsibility; TL = Positive Impact on Teacher Leadership; TP = Positive Impact on Teaching Practices; and YE = Years of Experience

Figure 3.1 ***SEM Model for SEP framework***

All latent variables were developed using the TALIS teacher survey items and each variable reflected the degree of teacher perception. The definition of key latent variables is included in Appendix B. Years of teaching experience was also included in this analysis as a control variable for the impact on professional capacity.

Statistical Software for Analysis

The software Mplus version 8.1 (Muthén & Muthén, 1998-2017) was used to conduct both multi-group CFA and the measurement invariance test. Mplus provides helpful functions to conduct statistical analysis with latent variables (Muthén & Muthén, 2018). As described in an earlier section of this chapter, teacher samples were nested in school samples. For analytical purposes, it is necessary to take account of such a clustered data structure. Otherwise, the variance and standard errors of the analysis would be underestimated (Geiser, 2012; OECD, 2014; Snijders & Bosker, 2011). In order to control for cluster effects, I used the Mplus "type is complex" with the "cluster" option.

Again, sampling weights were used to give participating countries equal impact and contribution to the estimation of model parameters. The maximum likelihood estimation (MLR) was used to conduct robust analysis against violations of normality as it was an iterative numerical integration procedure. The preliminary review of missing data showed that the data were missing at random (MAR). This meant that the probability of a missing observation did not depend on the true score of a person with regard to the variable of interest (OECD 2014). Missing data were imputed using the expectation-maximization (EM) algorithm on the total sample.

To assess the model fit of CFA and SEM, this study examined several key indices that have been widely used. These included the comparative fit index (CFI), the Tucker-Lewis index (TLI), standardized root mean square residual (SRMR), and root-mean-square-error of approximation (RMSEA). Following the standard research conventions, CFI $\geq$ .90, TLI $\geq$ .90, SRMR $\leq$. 10, and RMSEA $\leq$ .08 were considered as an acceptably

adequate model fit (Hu & Bentler, 1999; Kline, 2015; OECD, 2014; Schermelleh-Engel, Moosbrugger, & Müller, 2003; Steiger, 1990).

Methodological Limitations

This study has several methodological limitations. First, it uses survey data, and all surveys face some limitations, including sampling error, measurement error, and non-respondent error (Fowler Jr, 2013). These potential errors limit how much the study can be generalized to the overall population. Next, this study analyzes the factors that are included in the survey questionnaires. Thus, it may not account for all factors that may affect the impacts of teacher evaluation. Furthermore, the U.S. TALIS 2013 data did not meet the international participation rate standards as described above. Thus, it may introduce potential for bias in estimates using the weights in the U.S. TALIS data file (Strizek et al., 2014). Besides, the exploratory approach of this study may limit the ability to draw definitive causal relationships among variables, although the study purports to explore the relationship of the variables rather than to examine a particular hypothesis. Further studies that address these methodological limitations will help to advance the research on teacher evaluation.

Chapter 4 Teacher Evaluation Policy in Four Countries

This chapter presents a descriptive analysis of principal and teacher survey results to address the following questions:

RQ1. How are national teacher evaluation policies implemented at the local level in the United States, Finland, Korea, and Japan?

Specifically, the following questions are investigated: *To what extent are teachers formally evaluated? Who evaluates teachers, and what evaluation methods are used? How do principals and teachers perceive the outcomes and impact of teacher evaluation and feedback?*

This chapter consists of three sections. First, this paper examines the survey results of teachers' perceptions of the effectiveness of teacher evaluation in their schools. This reveals the overall understanding of teachers concerning the role of teacher evaluation and feedback in their schools. The second section examines principals' understanding of teacher evaluation policy at the local level and within their schools. It investigates who conducts teacher evaluation and provides feedback, how the evaluation and feedback are implemented, and how they are perceived to impact teachers. Lastly, it compares teachers' and principals' perceptions of the outcomes of teacher evaluation – e.g. How they affect teacher accountability (rewards or sanctions) and/or opportunities for professional development. Each of these three categories of descriptive statistics is analyzed for each of the four countries. This descriptive exploration helps explain the similarities and differences of structural characteristics between the countries, as well as

the differences and similarities in the perception of the effectiveness of teacher evaluation policy.

A brief note on language is appropriate before undertaking the analysis. The TALIS surveys use the term "appraisal," which is defined as "when a teacher's work is reviewed by the principal, an external inspector, or by his or her colleagues" (OECD, 2014, p. 406). In the U.S., evaluation is a more commonly used word to describe the processes, indicators, and outcomes of determining a teacher's effectiveness. Because the TALIS surveys do not provide a distinction between the terms, they will be used interchangeably in this discussion, except where a distinction is appropriate. The term evaluation will be used more frequently.

Teacher Perception of the Overall Impact of Teacher Evaluation and Feedback

Teachers' general perceptions of how evaluation and feedback systems affect their teaching are indicated based on their responses to two survey items: "Teacher appraisal and feedback have little impact upon the way teachers teach in the classroom" and "Teacher appraisal and feedback are largely done to fulfill administrative requirements." The survey results, shown in Table 4.1, suggest that teachers perceive evaluation as primarily a routine administrative task rather than a mechanism that enhances teacher and student experiences: about 60 - 62 % of teachers in the U.S., Finland, and Korea agreed with the latter statement, and while this percentage was relatively lower in Japan, nearly half of teachers agreed. This indicates that a significant portion of teachers in all our countries perceived that evaluation and the feedback that they received were largely done to fulfill administrative requirements.

Thus, it is not surprising that teachers in these four countries often reported a limited impact of teacher evaluation and feedback on their teaching practices. About 32 – 50 % of lower secondary teachers reported that they either agreed or strongly agreed that the evaluation and feedback that they receive had little impact upon their classroom teaching. This response was particularly evident in Finland, where almost half (49.9%) of teachers agreed or strongly agreed with the statement. The percentages of teachers who found that teacher appraisal and feedback had little impact on their teaching were 41% in Korea, 40% in the U.S., and 32% in Japan.

Table 4.1 ***Teachers' Perception of Teacher Appraisal and Feedback Systems in Schools***

	Teacher appraisal and feedback have little impact upon the way teachers teach in the classroom		Teacher appraisal and feedback are largely done to fulfil administrative requirements	
	% agree or strongly agree	(S.E.)	% agree or strongly agree	(S.E.)
Finland	49.9	(1.0)	62.0	(1.3)
Japan	32.4	(1.0)	47.3	(1.1)
Korea	40.6	(1.0)	59.8	(1.2)
United States	39.4	(1.5)	60.1	(1.6)

Data derived from the teacher questionnaire (question 31).
Source: OECD, TALIS 2013 Database.
Note. Adapted from OECD (2014, 04, 25). *The OECD Teaching and Learning International Survey (TALIS) 2013 Results - Excel Figures and Tables*. OECD. http://www.oecd.org/education/school/talis-excel-figures-and-tables.htm

Teacher Evaluation Policy at the Local Level

The principal survey provides rich information for understanding the implementation of teacher evaluation policy at the local level as principals are directly involved in the development and/or implementation of school-level teacher evaluation policy. Also, they are, in most cases, the evaluators of teachers. To understand the formal teacher evaluation policy in four countries at the local level, the section analyzes

principal survey questions related to three questions: (1) To what extent are teachers formally evaluated in their schools? (2) Who is responsible for teacher evaluation?, and (3) What measures are used to evaluate teachers?

To What Extent are Teachers Formally Evaluated in Their Schools?

This section examines the percentage of lower secondary education teachers who have never been appraised by specific bodies or have never been appraised at all by their education system, as reported by school principals (Table 4.2). Except for Finland, all or most principals reported that all teachers were formally evaluated by specific bodies. In the U.S. and Korea, principals reported that all of their teachers were evaluated by one or more evaluators and, in Japan, only 3.8% of principals reported that their teachers had never been formally evaluated. Finland is unique: about 26% of principals answered that their teachers had never been formally evaluated by their education systems. This may reflect the Finnish school system and climate that cherishes the autonomy and professional discernment of teachers. Still, about three-quarters of Finnish principals reported that their teachers were formally evaluated in their schools.

In general, the survey results show that teacher evaluation is a widely implemented policy and practice, and schools in all four countries have developed some kind of teacher evaluation systems. While the scope of impact may be different across countries, it seems that most teachers can expect to be exposed to formal evaluation by a supervisor or other authorities.

Table 4.2 ***Percentage of Teachers Who Never Received Formal Appraisals (Principal Responses)***

	Generally never formally appraised	
	%	(S.E.)
Finland	25.9	(4.2)
Japan	3.8	(1.1)
Korea	0.0	(0.0)
United States	0.0	(0.0)

Data derived from the principal questionnaire (question 27).
Source: OECD, TALIS 2013 Database.
Note. Adapted from OECD (2014, 04, 25). *The OECD Teaching and Learning International Survey (TALIS) 2013 Results - Excel Figures and Tables*. OECD. http://www.oecd.org/education/school/talis-excel-figures-and-tables.htm

Who is Responsible for Teacher Evaluation?

While most teachers are evaluated, there is significant variation between the countries on the allocation of responsibility for evaluation (Table 4.3). While the responsibility of teacher evaluation was more likely to involve local administrators, peers, and responsible authorities outside the school in both Korea and Japan, in the U.S. and Finland, teachers were likely to be evaluated only by internal or locally based evaluators.

Table 4.3 ***Teachers Who Never Received Formal Appraisals by Specific Bodies (Principal Responses)***

	Never formally appraised by school administrators*		Never formally appraised by peers**		Never formally appraised by external individuals or bodies***	
	%	(S.E.)	%	(S.E.)	%	(S.E.)
Finland	28.4	(3.8)	88.1	(2.7)	77.6	(3.0)
Japan	5.4	(1.9)	34.3	(3.8)	30.0	(3.7)
Korea	0.9	(0.6)	5.2	(2.5)	42.9	(4.9)
United States	2.5.	(2.5)	34.6	(5.5)	74.5	(5.3)

Data derived from the principal questionnaire (question 27).
Source: OECD, TALIS 2013 Database.
*school principals and the school management team
**teacher's mentor or other teachers who are not part of the management team
*** Adapted from OECD (2014, 04, 25). *The OECD Teaching and Learning International Survey (TALIS) 2013 Results - Excel Figures and Tables*. OECD. http://www.oecd.org/education/school/talis-excel-figures-and-tables.htm

In Japan, about 70% of principals reported that their teachers had been evaluated by external individuals or bodies, and about 65% of principals reported had been evaluated by mentors or other peers who were not in the management team. It seems that, in Japan, the responsibility of formal teacher evaluation is balanced and distributed not only within schools but also between internal and external bodies of evaluators. In Korea, peer evaluation is prominent compared to other countries: Only 5.2% of principals reported that their teachers have never been evaluated by their mentors or colleague teachers. In addition, about 43% of principals indicated that some external bodies or individuals had participated in the teacher evaluation process. In contrast, the likelihood of teachers being evaluated by external bodies or individuals was much lower in both the U.S. and Finland. Formal appraisal by external groups and peers was particularly rare in Finland, where more than 88% of principals reported that their teachers had never been evaluated by peers, and 78% reported no evaluations by external bodies. In the U.S, about three-quarters of principals reported that external bodies had never evaluated their teachers. Peer review was more common, although principals also reported that about one-third of teachers had never been evaluated by their peers. This means that teachers in Korea and Japan tend to have more opportunities to be evaluated by multiple bodies or individuals, both internally and externally, while teachers in Finland and the U.S. tend to be evaluated mainly by school administrators.

Despite these differences in the level of shared responsibility for evaluation, it is common in all four countries that school principals or school management teams operate as primary evaluators of teachers. Only 0.9%, 2.5%, and 5.4% of teachers had never been

formally evaluated by their school administrators in Korea, the U.S., and Japan, respectively. The percentage of principals who reported their teachers have never been evaluated by school administrators is relatively high in Finland (28.4%), compared to Japan or Korea, but where evaluation occurs, the principal or other school administrators have the primary responsibility. This result is, of course, consistent with the policy of decentralizing teacher evaluation, discussed in Chapter 2.

What Evaluation Processes and Indicators are Used to Evaluate Teachers?

Performance assessment may use a variety of processes (surveys, observations, interviews) and varied indicators of performance (client/parent satisfaction, observed skills, impact measure). The evaluation processes and indicators selected by the OECD surveys are:

- Direct observation of classroom teaching
- Assessment of teachers' content knowledge
- Analysis of student test scores
- Discussion of teachers' self-assessments of their work
- Discussion about feedback received from parents or guardians
- Student surveys about teaching

Table 4.4 shows the principals' responses to the ways in which different processes and indicators were applied. Note that this table includes only the principals in each country who indicated that the teachers in their schools were formally appraised.

Table 4.4 ***Teacher Appraisal Methods (Principal Responses)***

	Direct observation of classroom teaching		Student surveys about teaching		Assessment of teachers' content knowledge		Analysis of student test scores		Discussion of teachers' self-assessments of their work		Discussion about feedback received from parents or guardians	
	%	(S.E.)	%	(S.E.)	%	(S.E.)	%	(S.E.)	%	(S.E.)	%	(S.E.)
Finland	78.3	(4.0)	85.3	(4.0)	37.8	(4.9)	73.8	(5.0)	60.1	(4.5)	97.9	(1.6)
Japan	98.4	(1.2)	86.5	(2.7)	63.6	(3.7)	97.6	(1.1)	92.1	(2.2)	86.8	(2.4)
Korea	100.0	(0.0)	93.8	(2.0)	82.2	(3.3)	98.7	(0.9)	79.9	(3.3)	81.4	(3.2)
U.S.	100.0	(0.0)	60.1	(5.7)	72.1	(5.2)	93.3	(3.8)	73.7	(5.5)	90.5	(3.2)

Data derived from the principal questionnaire (question 28).

Source: OECD, TALIS 2013 Database.

Note. Adapted from OECD (2014, 04, 25). *The OECD Teaching and Learning International Survey (TALIS) 2013 Results - Excel Figures and Tables*. OECD. http://www.oecd.org/education/school/talis-excel-figures-and-tables.htm

First, three teacher evaluation processes are widely used in all four countries. Those include the direct observation of classroom teaching, the analysis of student test scores, and discussions about feedback from parents or guardians. In the U.S., Korea, and Japan, about 93 % - 100% of principals reported that they used direct observation and student test scores as sources of teacher evaluation. In Finland, where teacher evaluation is less common, about three-quarters of principals reported that teachers are appraised by direct observation and student test scores. This is not, perhaps, surprising since these two measures have dominated the discussion of how teacher performance should be gauged in the policy and research circles. It is also worth noting that feedback received from parents and guardians, which is largely absent in national and international policy discussions, is frequently used in the teacher evaluation process at the local level. An overwhelming majority of principals reported that their teachers had been evaluated through discussions about feedback from parents or guardians: 99% in Finland, 91% in the U.S., 87% in Japan, and 81% in Korea.

Overall, other assessment processes, such as student surveys, the discussion of teacher self-assessments, and the analysis of teacher content knowledge, are also frequently used for evaluation. However, the range of use for these measures is more varied between the four countries, compared to the three measures discussed above. In Finland, only about 38% of principals indicated that assessment of content knowledge was used to evaluate teachers, while about 85% reported that teachers had been evaluated using student surveys, and 60% indicated the use of teacher self-assessment. In the U.S., Korea, and Japan, a majority of principals indicated that they used all three of those measures for teacher evaluation.

The survey results suggest that those teachers who are evaluated in each of the four countries are assessed using multiple indicators, although there is a wide range in how widely each of the indicators is used. In Finland, for example, "user satisfaction" indicators from student surveys (85%) and parental feedback (98%) were used more frequently than other teacher evaluation methods like direct observation (78%) or student test scores analysis (74%), which are not considered to be conventional teacher evaluation measures in U.S. policy discourse. In Korea, where teachers were likely to receive evaluations from multiple sources, around 80% or more of the principals indicated that every indicator was used. Japan's profile is similar, though scoring somewhat lower on content knowledge assessments. The U.S., where most evaluation is carried out by local administrators, was least likely to report using student surveys, but a virtual consensus (90-100%) used observation, test scores, and parental feedback.

Outcomes of Teacher Evaluation and Feedback System

Scholars argue that effective teacher evaluation aims to achieve the dual purposes of teacher accountability for student learning as well as professional development to improve teacher performance (Darling-Hammond, 2013; Hargreaves & Braun, 2013). To understand the aims of teacher evaluation in these four countries, this section examines the perception of teachers and principals regarding what outcomes occur as a result of teacher evaluation. Both the teacher and principal TALIS surveys contain questions related to the outcomes of teacher evaluation and feedback. While the principal survey focuses on formal teacher evaluation outcomes, the teacher survey examines the outcomes of broad teacher evaluation and feedback systems. The teacher survey explains that "appraisal" is defined as a review of teachers' work, ranging from a more formal approach (e.g., as part of a formal performance management system, involving set procedures and criteria) to a more informal approach (e.g., through informal discussions). By doing so, the teacher survey may capture teachers' daily experiences of evaluation and feedback more broadly than the principal survey.

Analyzing the results of both surveys provides information on schools' evaluation policies and feedback systems from multiple angles. The principal survey provides information on what policy levers are available at the local level through the formal teacher evaluation process. On the other hand, the teacher survey provides a picture of how teachers understand the practice, processes, and outcomes of teacher assessment. The analysis of two kinds of surveys helps explore not only the similar trends of the

understandings of teachers and principals in teacher evaluation but also the gap between the perceptions of these two groups.

Outcome measures fit into two broad categories – those that emphasize teacher accountability and those intend to enhance professional growth. Outcomes that highlight teacher accountability may encompass changes in working conditions, changes in financial rewards, dismissal, or public recognition. On the other hand, outcomes related to professional development may include measures, such as thorough feedback, efforts to remedy specific weaknesses, mentor assignment, training plan development, etc.

Outcomes Related to Teacher Accountability

This section first analyzes the overall trends related to accountability outcomes in each country as evidenced by both the principal and teacher surveys. Then, it delineates the differences between principal and teacher survey results to understand how teachers perceive the accountability outcomes that occur in practice.

Concerning accountability, three areas of outcomes that hold teachers accountable for their teaching are analyzed with five questions in principal surveys, including (a) Material sanctions imposed on poor performers (e.g., reduced annual increases in pay), (b) An increase in teachers' salary or a payment of a financial bonus, (c) A change in teachers' work responsibilities, (d) A change in the likelihood of career advancement, and (e) Dismissal or non-renewal of contract. These five questions are then categorized and analyzed in three categories, as follows:

- A change in financial outcomes (questions a+b);
- A change in ' work conditions (questions c+d);
- Dismissal or non-renewal of contract.

Table 4.5 ***Accountability Related Outcomes of Formal Teacher Appraisal (Principal Responses)***

	Accountability					
	A change in financial outcomes*		A change in teachers' work conditions**		Dismissal or non-renewal of contract	
	%	(S.E.)	%	(S.E.)	%	(S.E.)
Finland	63.2	(4.1)	83.8	(3.1)	70.3	(5.0)
Japan	17.6	(2.9)	53.3	(4.0)	9.0	(2.1)
Korea	53.3	(4.7)	97.3	(1.8)	23.2	(3.7)
U.S.	32.7	(5.3)	81.6	(4.8)	94.6	(2.1)

Percent reporting that outcomes occurred "sometimes", "most of the time" or "always" after formal teacher appraisal.

Please note that schools that are not using formal teacher appraisal are not included.

*Financial bonus, salary or material sanctions (e.g., reduced annual increases in pay)

** Job responsibilities or the likelihood of career advancement

*** Adapted from OECD (2014, 04, 25). *The OECD Teaching and Learning International Survey (TALIS) 2013 Results - Excel Figures and Tables*. OECD. http://www.oecd.org/education/school/talis-excel-figures-and-tables.htm

Source: OECD, TALIS 2013 Database.

Table 4.6 ***Accountability Related Outcomes of Teacher Appraisal And Feedback Systems In Schools (Teacher Responses)***

	Accountability			
	If a teacher is consistently underperforming, he/she would be dismissed		The best performing teachers in this school receive the greatest recognition	
	%	(S.E.)	%	(S.E.)
Finland	16.4	(1.0)	25.3	(1.3)
Japan	13.9	(0.9)	37.1	(1.1)
Korea	18.9	(1.0)	51.0	(1.2)
U.S.	46.9	(2.3)	40.8	(2.1)

Percentage of lower secondary education teachers who agree or strongly agree. Schools that are not using formal teacher appraisal are not included

Source: OECD, TALIS 2013 Database.

Note. Adapted from OECD (2014, 04, 25). *The OECD Teaching and Learning International Survey (TALIS) 2013 Results - Excel Figures and Tables*. OECD. http://www.oecd.org/education/school/talis-excel-figures-and-tables.htm

Table 4.5 shows that a larger percentage of principals in the U.S. and Finland reported a change in working conditions and/or dismissal/non-renewal of contract (ranging from 70-95%) compared to a change in financial outcomes (about 32% in the U.S. and 63% in Finland). Notably, the percentage of principals who reported that

dismissal or contract non-renewal occurred as an outcome of teacher evaluation was particularly high in the U.S., reaching 95%. U.S. teacher surveys also indicated that about half of teachers perceived that they or their colleagues could be dismissed based on evaluations (Table 4.6). In Finland, Korea, and Japan, only 14-19 % of teachers perceived that dismissal or contract non-renewal occurred as a result of teacher evaluation and feedback.

On the other hand, changes in financial outcomes or in working conditions seemed to accompany teacher evaluation more frequently in Korea and Japan (Table 4.5). About 97% of principals in Korea reported that teachers' working conditions changed based on performance evaluations, while only 23% of principals reported that teachers were dismissed or their contracts were not renewed. Similarly, in Japan, only 9% of principals indicated that dismissal or non-renewal of contract occurred as an evaluation outcome, while 53% indicated that teachers' working conditions were changed as a result of the evaluation.

The TALIS teacher survey asks if teachers observe positive (i.e., public recognition of good performance) and negative (i.e., dismissal) accountability outcomes as a result of the evaluation. Teachers in Finland, Korea, and Japan responded that positive outcomes occurred more frequently than negative ones (see Table 4.5). Fewer than 20% of teachers in those three countries expected that a teacher would be dismissed if he/she was consistently underperforming. In contrast, the percentage of U.S. teachers who reported that punitive accountability outcomes followed from evaluation (47%) were higher than that of teachers who responded that positive or rewarding accountability

outcomes occurred (41%). Overall, the percentage of teachers who perceived that dismissal or non-renewal of contracts would happen as a result of teacher evaluation was significantly lower than that of principals. This could partly be because the teacher survey questions encompassed evaluation in a broader sense, including both formal and informal facets. Alternatively, it could be that teachers perceived that accountability outcomes of teacher evaluation were rare in practice.

Outcomes Related to Professional Development

Like the previous section, this section analyzes overall trends in both principal and teacher surveys for each country and then delineates the differences between principal and teacher survey results. Three categories of outcomes related to professional development are analyzed using both teacher and principal surveys: (1) *discussion* of remedies any weaknesses in teaching with teachers, (2) Assignment of a *mentor* to help the teacher improve his/her teaching, and (3) the development of *a professional improvement plan* for each teacher. Additionally, the teacher survey has one more outcome measure, which is (4) providing feedback to teachers based on a *thorough assessment* of their teaching.

Overall, every outcome related to professional development occurred widely (reported by half or more principals) as a result of teacher evaluation in all four countries. The discussion of measures to remedy weaknesses was most commonly reported, with 98-100% of principals indicating that, after a formal teacher evaluation, they discussed remedies to address any weaknesses always, most of the time, or, at least, some of the time (Table 4.7). Table 4.8 shows that teachers also reported the discussion of ways to

improve as the most common outcome of evaluation in all four countries, ranging from 65-75%.

Table 4.7 ***Professional Growth Related Outcomes of Formal Teacher Appraisal (Principal Responses)***

	Development					
	Measures to remedy any weaknesses in teaching are discussed with the teacher		A mentor is appointed to help the teacher improve his/her teaching		A development or training plan is developed for each teacher	
	%	(S.E.)	%	(S.E.)	%	(S.E.)
Finland	100.0	(0.0)	48.3	(5.0)	65.3	(5.2)
Japan	98.3	(1.0)	44.5	(3.5)	83.4	(2.8)
Korea	99.4	(0.6)	91.1	(2.4)	100.0	(0.0)
U.S.	100.0	(0.0)	86.5	(4.0)	96.6	(2.5)

Percentage reporting "sometimes", "most of the time" or "always". Schools that are not using formal teacher appraisal are not included.
Source: OECD, TALIS 2013 Database.
Note. Adapted from OECD (2014, 04, 25). *The OECD Teaching and Learning International Survey (TALIS) 2013 Results - Excel Figures and Tables*. OECD. http://www.oecd.org/education/school/talis-excel-figures-and-tables.htm

Table 4.8 ***Professional Growth Related Outcomes of Teacher Appraisal and Feedback Systems in Schools (Teacher Responses)***

	Development							
	Measures to remedy any weaknesses in teaching are discussed with the teacher		A mentor is appointed to help teachers improve his/her teaching		A development or training plan is established to improve their work as a teacher		Feedback is provided to teachers based on a thorough assessment of their teaching	
	%	(S.E.)	%	(S.E.)	%	(S.E.)	%	(S.E.)
Finland	65.2	(1.2)	16.5	(1.3)	38.5	(1.5)	16.8	(0.8)
Japan	70.6	(0.9)	31.4	(1.2)	45.6	(1.2)	31.6	(1.1)
Korea	75.4	(1.0)	46.1	(1.3)	69.4	(1.1)	50.1	(1.2)
U.S.	70.8	(2.0)	53.3	(2.0)	56.6	(2.0)	53.2	(2.2)

Percentage who "agree" or "strongly agree:
Source: OECD, TALIS 2013 Database.
Note. Adapted from OECD (2014, 04, 25). *The OECD Teaching and Learning International Survey (TALIS) 2013 Results - Excel Figures and Tables*. OECD. http://www.oecd.org/education/school/talis-excel-figures-and-tables.htm

Based on principal survey data (table 4.7), the U.S. and Korea appear to have developed balanced systems in which all three outcomes related to professional

development occur widely. In Korea, 99% of principals reported that formal teacher evaluation was followed by a discussion of ways to remedy weakness; 91% reported mentor assignment; and 100% reported the development of a development/training plan although the frequency of each outcome might vary. In the U.S., 100 % of principals reported that evaluation was followed by a discussion of ways to remedy weakness; 87%, reported mentor assignment; and 97% reported the development of a development /training plan, though again, the frequency of each outcome might vary. In contrast, in Japan and Finland, mentor assignment or development/training plans as levers for improvement occurred much less often than the discussion of ways to remedy weaknesses: about 45% of principals in Japan and 48% of principals in Finland responded that they had appointed mentors to help teachers and about 48% in Japan and 65% in Finland reported that they had developed a development or training plan after evaluation.

Teachers and principals do not, however, see the outcomes in the same way. When comparing the principal survey results (Table 4.7) with teacher survey results (Table 4.8), the percentage of teachers who agreed or strongly agreed that these three development outcomes occurred was much lower than the percentage of principals who answered that these outcomes occurred at their schools in all four countries. In Finland, for example, only 17% of teachers perceived that a mentor would be assigned to help teacher's teaching, and 39% of teachers thought that development/training plans were established as a part of the teacher evaluation and feedback process. The discrepancy between principal and teacher surveys suggests that although there are multiple

administrative instruments to support teacher professional development through teacher evaluation, these may not be used widely in practice, or at least publicly enough for teachers to observe them as a school policy. In addition, overall, one may reasonably conclude that discussions and paper plans are more likely to occur in each country than the assignment of additional support through a mentor or more analysis of their teaching practice.

The teacher survey also shows a possible link between teachers' perception of the extent to which they can expect to receive feedback through the rigorous assessment of their teaching and the extent to which they might be assigned a mentor. For instance, in the U.S., about 53% of teachers agreed or strongly agreed that a mentor would be appointed to help teachers, and the same percentage of teachers perceived that feedback was provided based on a thorough assessment of their teaching. In Finland, about 17% of teachers agreed or strongly agreed that a mentor would be appointed to help teachers and that feedback was provided based on a thorough assessment of their teaching. This trend was also observed in Korea and Japan.

Discussion of Descriptive Analysis

The surveys reveal some general trends in all four countries as well as several unique features of teacher evaluation systems in each country.

Views on the Effectiveness of Teacher Evaluation

Overall, there is a gap between the teacher and principal survey results regarding their views on teacher evaluation policy implementation. The percentages of teachers who responded that certain outcomes occurred as a result of teacher evaluation and

feedback that they received were significantly lower than the percentages of principals who responded that corresponding outcomes were followed after evaluation. This might mean that teachers perceived that evaluation and feedback were not connected to meaningful outcomes in practice. This was consistent with the survey result indicating that a large portion of teachers believed that teacher evaluation and feedback were done largely to fulfill administrative requirements. A significant portion of teachers in all four countries perceived that teacher appraisal and feedback had little impact upon classroom teaching (i.e., 49.9% in Finland, 41% in Korea, 40% in the U.S., and 32 % in Japan).

These survey results raise questions about the effectiveness of teacher evaluation and feedback that teachers received at their schools, particularly considering the impact of teacher evaluation on classroom teaching and student learning. Teacher evaluation policy was widely implemented at the local level in all countries but appears to fall short as an effective and meaningful tool to enhance teaching practices and student learning in classrooms.

Characteristics of Teacher Evaluation in Four Countries

Although there were trends commonly observed in all four countries, differences also appeared. Out of all four countries, Korea seemed to have the most balanced teacher evaluation system. The responsibility of teacher evaluation was shared among multiple evaluators, both internally and externally; multiple evaluation processes and indicators were used to evaluate teachers. Also, multiple administrative policy instruments were available to shape teacher evaluation outcomes. In the U.S., teacher evaluation was centered in the school, where teachers were assessed primarily by principals but quite

often by peers. In contrast, the percentage of teachers who have been evaluated by any external group was markedly low (about 25%). In Japan, the school principal or management team was also the main evaluator of teachers, but the involvement of peers and external groups was also quite high (75% and 70% respectively).

Finland was unique among the three countries in this study: Teachers were far less likely to be evaluated by local administrators, peers, or their school administrator, and about a quarter of principals reported that their teachers have never been formally appraised. Among those who did report teacher appraisal in their schools, surveys from students and feedback from parents were more widely used than direct observation or teacher self-assessment/content knowledge assessment. Finnish teachers were the most likely to say that formal evaluation had little or no impact on their teaching or classrooms. Rather than an evaluation system, the Finnish approach to teacher assessment appeared to be driven by collaboration and feedback between individual teachers and the students and parents with whom they work, with some guidance provided (in some cases) by discussions with the school administrator.

Although global policy debates on teacher evaluation have largely been centered on evaluation as a lever to hold teachers accountable for student learning, the percentage of teachers who reported accountability related outcomes, such as dismissal or changes in working conditions, was low compared to those who perceived professional development outcomes following evaluation. This trend was particularly apparent in the three high-performing countries. In Korea and Japan, the more elaborated evaluation systems appeared to shape teacher quality by changing their working conditions or financial

rewards, rather than by dismissal. In Finland, assessment appeared to be a weaker lever for either dismissal or rewards. Only in the U.S., where about half of teachers believed that dismissal and non-renewal of contract might occur as a result of teacher evaluation, was the evaluation system viewed primarily as a potential hammer. Yet, even in the U.S., the use of multiple within-school instruments (mentors, improvement plans, and more rigorous analysis of teaching) as well as recognition of good teaching, were widely used than high-stakes instruments.

Chapter 5 Implementing Effective Teacher Evaluation

This chapter explores the relationships between teacher evaluation policy and other organizational factors. Two research questions are addressed:

RQ2. How are teacher evaluation policies and school climate associated with teachers' perceived professional capacity in four countries?

RQ3. How are teacher evaluations associated with teachers' professional capacities when evaluation is accompanied by support of teacher professional growth?

As implied by the research questions, the analysis proceeded in two parts. First, it analyzed the association among the organizational elements of effective teacher evaluation policy within the four countries using the Structural Equation Modeling (SEM). It particularly focused on the relationships among the outcomes of teacher evaluation and feedback, the school climate of shared responsibility, and the perceived positive impact of teacher evaluation and feedback. More specifically, SEM examined how the teacher evaluation outcomes for professional growth and the school climate of shared responsibility were associated with professional capacity.

Second, SEM was further used to examine the extent to which teacher evaluation, when it was closely connected to the professional development outcomes, was associated with the school improvement factors: (1) teaching practices, (2) teacher motivation/job satisfaction/confidence, and (3) teacher leadership. By analyzing the relationships among teacher evaluation outcomes, the impact of teacher evaluation and feedback, and the

school climate of shared responsibility, the relationships among teacher evaluation and multiple teacher-related factors in conjunction with school climate were described.

Defining Organizational Variables for the Structural Equation Models

A multi-group confirmatory factor analysis was conducted to generate latent variables across all four countries. An investigation on the normality of the dataset was conducted using measures of skewness and kurtosis (see Appendix C). As the results indicated that the data did not meet the normality assumptions for a few variables, bootstrapping was used that helped correct for these biases (Hesterberg, 2011; Kline, 2015).

Multi-Group Confirmatory Factor Analysis

Multi-group Confirmatory Factor Analysis (CFA) was carried out to define and measure the impacts of teacher evaluation outcomes in supporting teacher professional growth and development in the four nations. Three latent variables emerged from the 22 indicators shown in Table 5.1. A positive impact on professional capacity was proposed as a second-order factor that consisted of three subscales: *motivation/confidence/job satisfaction, teaching practices,* and *teacher leadership*. It encompassed various aspects of daily classroom practice in addition to emotional and social satisfaction with teaching. The scale of *shared responsibility* represented school climate and contextual elements that promoted shared responsibility. It is noteworthy that *teacher evaluation and feedback outcomes that were linked to professional growth* was validated as a factor. In contrast, teacher evaluation outcomes linked to high-stakes results (such as dismissal or financial

bonuses) did not emerge as a validated factor across the four countries and were, therefore, eliminated from further analysis.

Table 5.1 ***Latent Variables and Survey Items***

Scale	Subscale	Variable	Item
Evaluation and feedback outcomes for professional growth		TT2G31D	Development/training plan
		TT2G31E	Feedback from thorough evaluation
		TT2G31G	Weakness remedy measures
		TT2G31H	Mentor assignment
Shared responsibility		TT2G44A	Staff participation in school decisions
		TT2G44B	Parents/guardians participation in school decisions
		TT2G44C	Students participation in school decisions
		TT2G44D	Culture of shared responsibility for school issues
		TT2G44E	Collaborative culture with mutual support
Perceived positive impact on professional capacity	Motivation, Confidence, Satisfaction	TT2G30F	Confidence as a teacher
		TT2G30M	Job satisfaction
		TT2G30N	Motivation
	Teaching practices	TT2G30H	Classroom management practices
		TT2G30I	Knowledge and understanding of your main subject field(s)
		TT2G30J	Teaching practices
		TT2G30K	Methods for teaching students with special needs
		TT2G30L	Use of student assessments to improve student learning
	Teacher leadership	TT2G30A	Public recognition from the principal and/or your colleagues
		TT2G30B	Role in school development initiatives
		TT2G30C	Career advancement likelihood
		TT2G30D	Amount of professional development
		TT2G30E	Job responsibilities at this school

The multi-group CFA indicated an acceptable or good overall model fit in all four countries (Geiser, 2012; Kline, 2015; OECD, 2014): for each country, the CFI was higher than .9, the SRMR was less than .05, and the RMSEA was less than 1.0 (see Table 5.2). Especially, with the U.S., the model fit measures indicated a good fit, with the CFI and TLI each being higher or equal to .95. With Japan, the TLI was a bit lower than .9; the rest of the model fit measures indicated an acceptable fit for this country.

Table 5.2 ***Multi-Group Confirmatory Factor Analysis Model Fit for All Four Countries***

Country	CFI	TLI	RMSEA	SRMR
U.S.	.957	.950	.039	.035
Finland	.930	.920	.041	.044
Korean	.933	.922	.055	.031
Japan	.900	.884	.057	.046

To further validate the three-factor model, a competing model was tested that included one first-order latent variable (*i.e., overall perceived positive impact*) that encompassed all the survey items on the positive impact of the teacher evaluation and feedback system in the schools. Sub-scales of positive impact on teaching practices, teacher motivation/job-satisfaction/confidence, and teacher leadership were not separated. The alternative model was developed based on the research that suggested a comprehensive impact of effective teacher evaluation on teaching and teachers (Darling-Hammond, 2013). The results indicated that the fit of the proposed three-factor model was better (see Appendix D). Also, the proposed model allowed the analysis of the association between the teacher evaluation policy and its impact on three different aspects of teaching.

Based on the three-factor model developed, the factor loadings were investigated to test their psychometric properties and construct validity. Tabachnick, Fidell, and Ullman (2007) suggest that an excellent factor loading is higher than .70, and a good factor loading is higher than .50. All the indicator variables in the individual U.S., Finland, and Korea models and most of the indicator variables in the Japan model showed excellent or good factor loadings. Although the factor loading of the teacher evaluation outcomes to the implementation of measures to remedy weakness (.450) fell below the recommended standards and was were relatively lower than other indicators, their coefficients for factor loadings were also statistically significant *($p < .001$)*. The analysis therefore proceeded with its inclusion.

Validating Between Country Comparisons: Measurement Invariance Test

Measurement invariance testing is often used in international comparative studies to determine if an equivalent construct is being measured across different countries (Chen, 2007; Kline, 2015; OECD, 2014; Putnick & Bornstein, 2016). Measurement invariance indicates whether a specified measure is construed in a conceptually similar manner by survey participants that represent different groups or cultural backgrounds. Thus, the measurement invariance test ensures the meaningful interpretation of cross-national measurement data.

In this study, the invariance testing was conducted through the multi-group CFA. Following the precedence of the Teaching and Learning International Survey (TALIS) technical report (2013), three levels of invariance were examined, which included configural, metric, and scalar levels of invariance. Configural invariance is achieved when the factor structures are equivalent across all countries in which the same variables are associated with each of the common factors. Metric invariance is achieved when the same dimensional structure is found, and the magnitudes of the associations between the variables and the corresponding factor are equal across the different countries. Metric invariance means the factor loadings are similar across countries. Scalar invariance is achieved when the intercepts observed in each country are equivalent. This shows that the values and means are also the same across the groups.

Table 5.3 presents the results of the measurement invariance test applied to the three latent factors for the four countries. First, the comparison between the unrestricted multiple-group model (i.e., configural invariance) and the model with equal factor

loadings across countries (i.e., metric invariance) was conducted to confirm validity for cross-cultural comparisons of the correlations of these factors with other constructs.

The criteria of a .01 absolute difference in CFI and .01 difference in RMSEA are commonly used, although some researchers have suggested other criteria for the CFI or the use of alternative fit indices, such as ΔRMSEA or ΔSRMR. Chen (2007) suggested a cutoff of a .01 absolute difference in the CFI (ΔCFI ≤ .01), paired with changes in the RMSEA of .015 (ΔRMSEA ≤ .015) and an SRMR of .030 (ΔSRMR ≤ .030) for a metric invariance or .015 (ΔSRMR ≤ .030) for scalar or residual invariance when the sample size is adequate (total N > 300) and the sample sizes are equal across the groups when the lack of invariance is mixed. Chen (2007) chose the CFI as the main criterion given that the RMSEA and SRMR tended to over reject an invariant model when the sample size was small. Rutkowski and Svetina (2014) suggested using a more lenient criterion when comparing 10 or 20 groups. They concluded that changes in the CFI of less than .02 and changes in the RMSEA of less than .03 were most appropriate for tests of metric invariance with large group sizes. Following the widely used criteria in the measurement invariance test, this study examined the differences in the CFI and RMSEA (ΔCFI ≤ .01 & ΔRMSEA ≤ .01) while also attending to the change in the SRMR.

The results showed that the model data fit of the configural invariance was good. The differences between the configural and metric levels of invariance were small with first-order factors and with both first- and second-order factors. The differences of the SRMR among the two levels were all tenable, and the differences of the CFI were in the acceptable range. However, the difference between the scalar (i.e., the model with equal

factor loadings and item intercepts) and metric levels of invariance was considered somewhat outside the acceptable range. Thus, analyzing the mean score comparisons across countries should be conducted with great care because a mean score may have a slightly different meaning in each country (as was noted in Chapter 3) as a limitation of the OECD survey. In sum, the analysis confirmed the validity for the cross-cultural comparisons of the correlations of these latent factors with other constructs. However, it also implies that the mean score comparisons for the scale cannot be explicitly interpreted.

Table 5.3 ***Measurement Invariance Tests of the First- and Second-Order Latent Factors***

Invariance Level	CFI	TLI	RMSEA	SRMR	ΔCFI	ΔRMSEA	ΔSRMR
Configural invariance	.966	.962	.035	.020	-	-	-
Metric invariance of the first-order factors	.964	.960	.035	.028	.002	.000	.008
Metric invariance of the first & second-order factors	.963	.960	.035	.028	.003	.000	.008
Scalar invariance of the first-order factors	.946	.944	.042	.038	.018	.007	.010
Scalar invariance of the first & second-order factors	.935	.935	.045	.073	.028	.008	.045

SEM Analysis I: Effective Teacher Evaluation and Organizational Contexts

After validating the constructs for the cross-national analysis, SEM was conducted to understand how the *evaluation and feedback outcomes for professional growth* (EFP) were associated with school climate (i.e., *shared responsibility*, or SR) and *professional capacity* (PC). As shown in Table 5.1, the teacher evaluation and feedback variable consisted of four survey items in which teachers rated the outcomes of teacher evaluation and feedback. All the items contributing to the teacher evaluation and feedback outcomes variable were indicators of support for teacher professional growth (i.e., development/training plan, feedback from thorough evaluation, weakness remedy

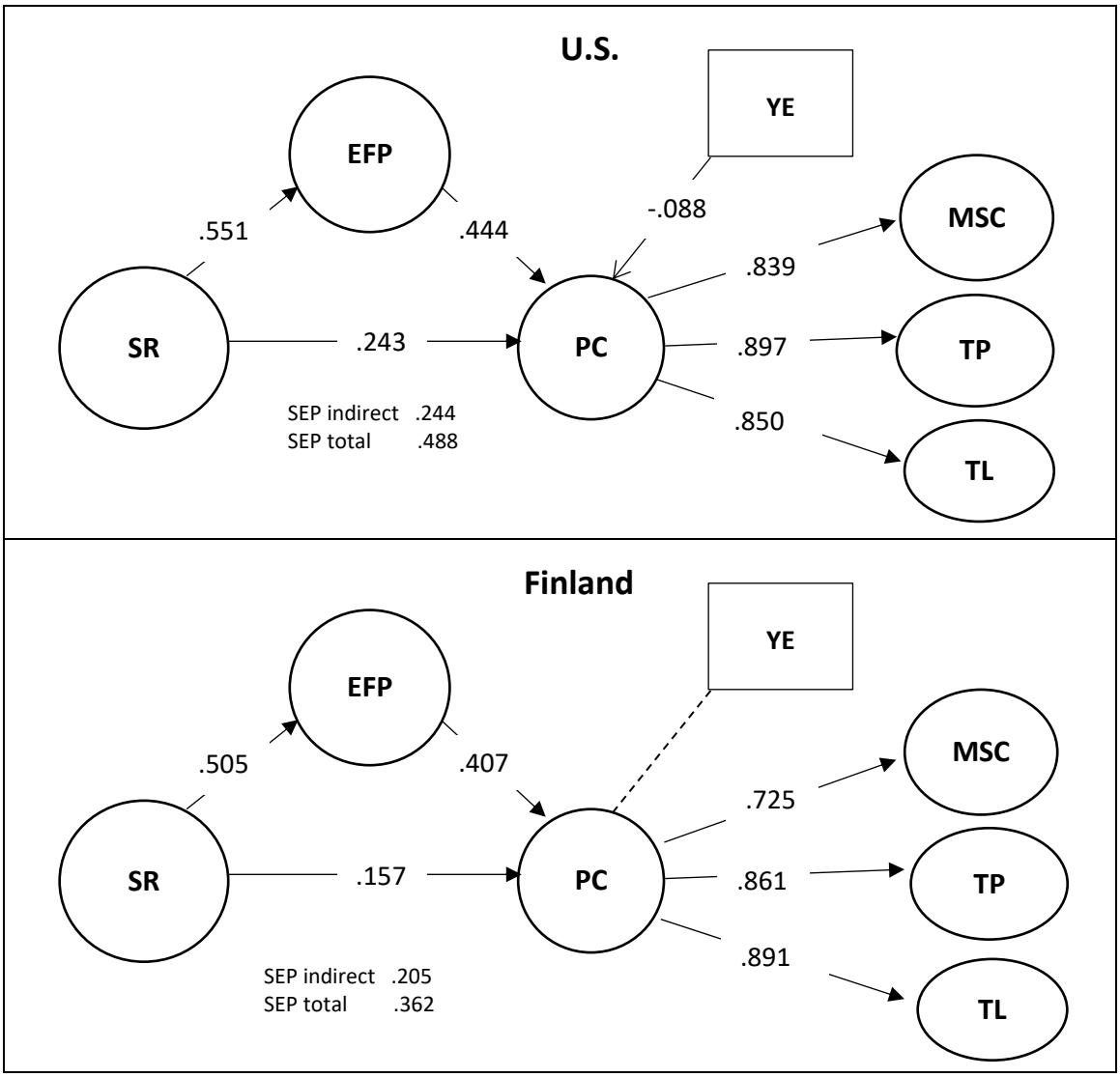
measures, and mentor assignment). In other words, the *evaluation and feedback outcomes for professional growth* (EFP) variable showed to what extent teacher evaluation was implemented in a way that supports teacher development. The *total years of teaching experience* (YE) was used as a control variable in the analysis. To understand the *shared responsibility* (SR) effects on EFP and *professional capacity* (PC) in the four countries, the samples of the four countries were analyzed using the same factors consistently, and the results were compared. Those samples were analyzed using the survey weights provided by the OECD while controlling for the cluster effects at the school level.

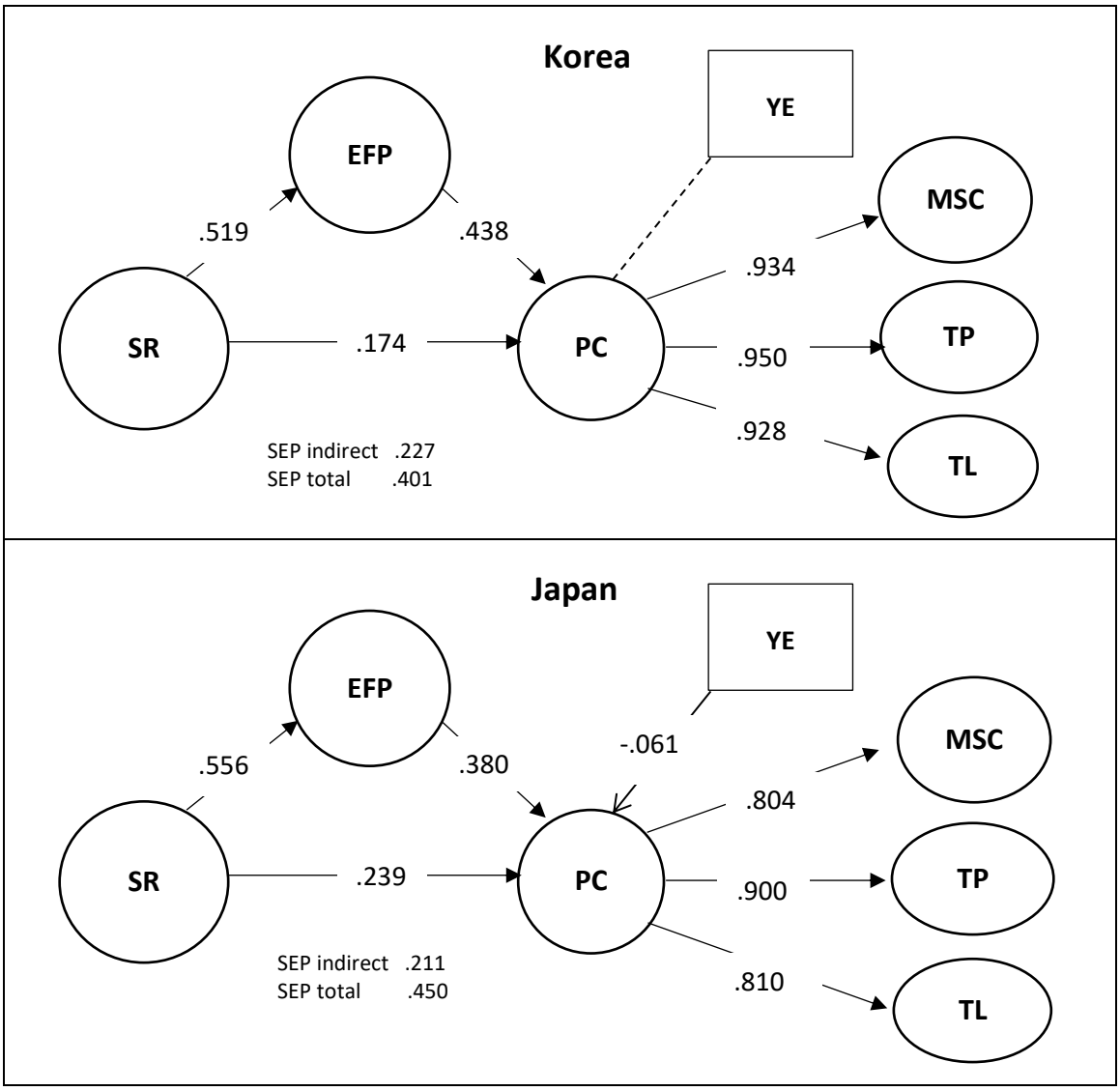
The SEM model showed an acceptable or good overall model fit in all four countries, although the fit was marginal with the Japan model (see Geiser, 2012; Kline, 2015; OECD, 2014). The model fit indices are included in Appendix F. Generally, similar patterns were consistently found across all four countries of this study. This result confirmed that first, teacher evaluation and feedback outcomes for professional growth was an important factor that had a moderate direct effect on teachers' perceived professional capacity. When the outcomes of teacher evaluation and feedback were linked to professional growth measures, teachers were more likely to perceive that it had a positive impact on their professional capacity. Second, the school climate in which the teacher evaluation was implemented was also critical. A school climate characterized by shared responsibility had moderate indirect effects on professional capacity through teacher evaluation and a small direct effect on professional capacity. Finally, the professional capacity factor encompassed various aspects of teaching. The general results

are summarized here, and a more detailed discussion of these findings is found in the following sections.

Direct and Indirect Effects of Shared Responsibility Through Teacher Evaluation

Figure 5.1 depicts the SEM results for each country based on the general model outlined in Chapter 3. In Figure 5.1, a significant effect is represented with an arrow, and a non-significant relationship is represented with a dotted line.





Note. EFP = Evaluation and Feedback Outcomes for Professional Growth; MSC = Positive Impact on Motivation, Job Satisfaction, and Confidence; PC = Professional Capacity; SR = Shared Responsibility; TL = Positive Impact on Teacher Leadership; TP = Positive Impact on Teaching Practices; and YE = Years of Experience

Figure 5.1 ***SEM Results of the SEP Model in the Four Countries***

In the U.S. model, the direct effect of *evaluation and feedback outcomes for professional growth* (EFP) on *professional capacity* (PC) was .444, and the direct effect of *shared responsibility* (SR) on *evaluation and feedback outcomes for professional growth* (EFP) was .551. Both effects were significant. When teacher evaluation and

feedback were linked to professional growth measures, they had a significant effect on teachers' perceived positive impact. When examining the indirect effect of *shared responsibility* (SR) in the U.S. model, the total effect of SR on PC was .488, and its indirect effects on PC through the EFP was .244. About fifty percent of the total effect (.244/488) was the indirect effect through the teacher evaluation and feedback system. This may indicate the importance of the use of teacher evaluation and feedback outcomes in the U.S. context.

In Finland, the direct effect of *evaluation and feedback outcomes for professional growth* (EFP) on *professional capacity* (PC) and the direct effect of *shared responsibility* (SR) on *evaluation and feedback outcomes for professional growth* (EFP) were significant, .407 and .505, respectively. Both the total and indirect effects of the SEP were significant. The examination of the SEP link indicated that about 57% of the total impact (.205/.362) between the shared responsibility and the professional capacity was the indirect effect through the teacher evaluation and feedback system. Note that the total effect of the SEP link in the Finland model was the lowest when compared to countries like Japan and Korea. These results indicated that the effect of shared responsibility on the perceived positive impact of teacher evaluation and feedback was relatively low compared to the other three countries.

The pattern in Korea's model was similar to that of Finland and the U.S. models. The direct effect of *evaluation and feedback outcomes for professional growth* (EFP) on *professional capacity* (PC) and the direct effect of *shared responsibility* (SR) on *evaluation and feedback outcomes for professional growth* (EFP) were significant, .438

and .519, respectively. Again, in Korea, teachers perceived that the shared responsibility, the teacher evaluation and feedback system, and the perceived positive impact of teacher evaluations were all associated with each other. Both the total and indirect effects of the SEP link were significant in the Korea model, and about 57% of the total impact (.227/.401) was the indirect effect through the teacher evaluation and feedback system.

The direct effect of *evaluation and feedback outcomes for professional growth* (EFP) on *professional capacity* (PC) in Japan was significant, but the effect size (.380) was slightly lower than that found in the other models of the other countries. The effects of *shared responsibility* (SR) on EFP and PC were both significant and relatively high compared to the other nations. This might show the importance of school climate when it comes to effective teacher evaluation policies and practices from the perspective of teachers in Japan. However, teacher evaluation still played a role in teacher professional capacity. About 47% of the total effect of the SEP link was an indirect effect through the teacher evaluation and feedback outcomes.

The review of SEM results showed that both shared responsibility and teacher evaluation and feedback outcomes for professional capacity were positively associated with the teachers' perceived impact on their professional capacity across all four country models. The analysis of indirect effect also implied that the effect of shared responsibility on professional capacity was influenced by how teacher evaluation outcomes were used to support teacher professional growth.

Direct Effects of Shared Responsibility on Teacher Evaluation

The previous section examined the association among three factors: *shared responsibility* (SR), *evaluation and feedback outcomes for professional growth* (EFP), and *professional capacity* (PC). This section further analyzes the relationship between the school climate and teacher evaluation implementation.

Figure 5.1 presents the direct effect of *shared responsibility* (SR) on *evaluation and feedback outcomes for professional growth* (EFP) in each country as significant. The effect sizes of SR on EFP were .551, .505, .519, and .556 in the U.S., Finland, Korea, and Japan, respectively. This indicated that the school climate of shared responsibility was moderately associated with the ways in which teacher evaluation and feedback were implemented and the outcomes that followed as a result of the teacher evaluation. In other words, when school leaders and staff created the school climate in which staff members, parents, and students participated in school decisions and collaborative culture on school issues occurred with mutual support, it was more likely that teacher evaluation outcomes for professional growth occurred—for instance, mentor assignments, ample feedback, and developing plans to remedy a weakness.

The effects of *shared responsibility* (SR) and *evaluation and feedback outcomes for professional growth* (EFP) on *professional capacity* (PC) also overwhelmed the effects of the teachers' years of experience. Across all four countries, the association of *years of experience* (YE) with perceived positive impact was either insignificant or minimal. This contradicts both previous research findings (discussed in Chapter 2) and conventional wisdom, which assumes that experienced teachers are less likely to change

their teaching practices as a result of new policies or reform initiatives (also discussed in Chapter 2). The results indicated that years of experience had no statistically significant relationship with the level of teachers' perception of the positive impact of EFP on PC in Finland or Korea. In both the U.S. and Japan models, there was a slightly negative significant effect of YE on PC. Nevertheless, the effect sizes in both cases were very small, which may indicate that they were of limited practical importance. This suggests that from a teacher's viewpoint, the outcomes of the teacher evaluation and feedback were associated with professional capacity regardless of their years of experience when the outcomes were implemented in a way that bolstered teacher professional development and growth.

SEM Analysis II: Relationships with Professional Capacity

From a policy perspective, the findings are of interest primarily in the context of the answer to the final question: *Does the perceived impact of teacher evaluation policies have an association with changes in teacher quality?* To answer this question, this section examines three components of professional capacity. The second-order latent variable of professional capacity consisted of three subscales of *teaching practices*, *teacher motivation/job satisfaction/confidence*, and *teacher leadership* in the analytical model. Each of these reflects capacities that are essential for teachers in developing their professional identity, competencies, and performance, and all have been positively associated with student learning growth and school improvement in prior research (see Chapter 2). Thus, it is meaningful to analyze how these three subscales comprise the

psychometric property of the professional capacity variable by examining their factor loadings.

Factor loadings indicate how much a factor explains each variable. Examining a factor loading helps in understanding how much the factor influences each variable. Loadings close to -1 or 1 indicate that the factor strongly influences the variable. Loadings close to zero indicate that the factor has a weak influence on the variable. Tabachnick et al. (2007) suggest that an excellent factor loading is higher than .70, and a good factor loading is higher than .50.

Table 5.4 ***Factor Loadings of Professional Capacity***

Scale	Subscale	Factor loading			
		U.S.	Finland	Korea	Japan
Professional Capacity	Motivation/Job satisfaction/Confidence	.839	.725	.934	.804
	Teaching practices	.897	.861	.950	.900
	Teacher leadership	.850	.891	.928	.810

Table 5.4 shows that all the subscales had excellent factor loadings in all the country models. This means that each subscale explained the latent variable of the professional capacity well in all four country models, and the latent variable influenced each variable strongly. In the Korea model, all three subscales had strong factor loadings of over .9, with the highest factor loading for the teaching practices subscale at .950. This indicates that professional capacity was well explained by each of the three subscales. In the U.S. and Japan, all three subscales had factor loadings over .8, with the highest factor loadings for the teaching practices subscales of .897 and .900, respectively. In Finland, the factor loading of motivation/job satisfaction/confidence was relatively lower at .725, while the factor loadings of both teaching practices and teacher leadership were over .8.

While there were some differences between the countries, they were slight in comparison to the consistency. This result confirmed the holistic and comprehensive definition of professional capacity that this dissertation has suggested.

Extending the analysis to the relationships between teacher evaluation and feedback outcomes and the professional capacities discussed above (see Figure 5.1), evaluation feedback outcomes that focused on professional development (EFP) had a statistically significant effect on teachers' professional capacity (PC), which, as discussed, comprised teacher motivation/job satisfaction/confidence, teaching practices, and teacher leadership. This implied that, when teacher evaluation and feedback were linked to teacher professional growth, their relationships with teachers could encompass daily classroom practice, emotional and social satisfaction with teaching (i.e., teacher motivation, job satisfaction, and confidence), and teachers' willingness to step into leadership activities.

Chapter 6 Discussion

Teacher quality is traditionally viewed as an individual trait that is developed through policies that promote rigorous initial selection, high-quality professional preparation programs, and effective hiring and retention practices (Executive Office of the President, 2002). Many recent studies, however, have emphasized the influence of school and district contexts on teaching practice (e.g., Kennedy, 2008; Knight et al., 2015), which suggests a need to examine how local policies, interactions among teachers, and teacher-administrator relationships influence the development of teachers' classroom practice. Thus, an organizational approach that was outlined in Chapter 1 and presented in greater detail in Chapter 2 begins with the assumption that educational system change, through national policies, can support quality teaching by improving local practices to support teacher improvement.

However, as was shown in Chapter 2, in many countries, efforts to develop a quality teaching force have also prioritized the idea of improving the evaluation of teachers, focusing particularly on the effectiveness of a teacher's practice in increasing student knowledge. Because teacher evaluation policies have become prominent around the globe, this study set out to examine three questions:

(1) How are national teacher evaluation policies implemented at the local level in four countries?

(2) How are teacher evaluation policies and school climate associated with teachers' perceived professional capacity in four countries?

(3) How are teacher evaluations associated with teachers' professional capacities when evaluation is accompanied by support of teacher professional growth?

Using an organizational approach, this study examined how teacher evaluation was implemented, how teachers' assessments of evaluation implementation were associated with school climate, and how school climate then affects teacher professional capacity. The analysis was conducted using data from four countries, the U.S. and three other educational systems that are widely regarded as among the best in producing high student achievement: Finland, Korea, and Japan. Synthesizing the results from both Chapters 4 and 5, this chapter summarizes what has been learned about teacher evaluation policy practices in these different countries, key factors for successful teacher evaluation implementation, and the potential implications for policymakers and school leaders.

Assessment of Teacher Evaluation in Four Countries

This study attempted to advance the notion of teacher evaluation research by focusing on the implementation of teacher evaluation policies in multiple countries. While teacher evaluation policies can be very complex, this study was centered on how teacher evaluation was implemented in ways that balanced both accountability and professional growth. A growing number of researchers highlighted that teacher evaluation needs to achieve these dual goals in order to improve teaching practices while also retaining good teachers (Darling-Hammond, 2013; Hargreaves & Braun, 2013; Smylie, 2014; Taylor & Tyler, 2012a).

To understand how teacher evaluation was implemented to achieve the goals of accountability and professional growth, policy narratives at the national level were examined along with the results of teacher/school leader surveys in schools, where, according to many scholars (e.g., Honig, 2006; Spillane, et al., 2002), the implementation of national/state/district teacher policies become entwined with school-level policies and practices. The analysis suggested that distinctive perspectives were developed at the national, school, and teacher levels.

For the last two decades, the four countries included in this analysis have launched national policy initiatives to overhaul their respective teacher evaluation policies. Although their approaches were different, they all aimed to improve teacher and teaching quality, which were expected to result in better student learning. Within this group, the U.S. pursued a path of accountability-driven teacher evaluation reform that began in the early 2000s. As seen in federal initiatives, such as the Highly Qualified Teacher (HQT) mandate in the No Child Left Behind (NCLB) Act of 2001 and federal actions during the Obama administration, including funding for the Race to The Top (RTT) program in 2009, policymakers conceived plans to hold individual educators more accountable for student performance. These federal and related state regulations initiated significant shifts in teacher evaluation in schools: they increased individual in-service teachers' accountability for student learning and expanded the federal regulations on specific ways to evaluate teachers. In many cases, the federal legislation tracked state initiatives that preceded them and merely added additional incentives and requirements.

This trend continued until the federal Every Student Succeeds Act (ESSA), which in 2015 loosened the accountability requirements as described in Chapter 2.

In the 2013 OECD U.S. TALIS survey sample, high percentages of both teachers and principals reported that the accountability measures could occur as a result of teacher evaluation (for more details, see Tables 4.5 and 4.6). For instance, about half of the teachers agreed or strongly agreed that if a teacher was consistently underperforming, he/she would be dismissed. In the other three countries, the percentage of teachers who agreed or strongly agreed with the above statement was less than 20%. However, the percentage of teachers who agreed or strongly agreed that professional growth outcomes happened was comparable to or higher than the accountability outcome percentages in all four countries.

During the same period, Korea and Japan both implemented the same two initiatives to revamp their teacher evaluation policies—performance-based pay and teacher evaluation for professional development. These two policy changes were passed as separate initiatives but were executed almost simultaneously in both countries around 2010, influencing their respective school practices. Policymakers and education leaders in Korea and Japan were explicit about the intent of the evaluation policies to achieve both accountability and professional development. In 2013, the TALIS teacher survey reported that multiple professional growth measures occurred after teacher evaluation in both Korea and Japan. Especially in Korea, the figures in all areas of the professional growth outcomes were higher in Korea compared to the other three countries.

However, both the Korean and Japanese efforts to implement multiple teacher professional growth outcomes through teacher evaluation seemed to be a partial success. Although multiple professional growth instruments were implemented, they were implemented in a climate that increased the accountability and pressure on individual teachers rather than prompting shared responsibility. Performance-based pay that focused on individual accountability and teacher evaluation for professional development did not appear to be implemented in complete harmony. Thus, these two initiatives caused confusion and frustration in the field, as they were seen as sending conflicting messages and were implemented in a policy climate that focused on teacher accountability, as described in Chapter 2. In Japan, teachers perceived that the new teacher evaluation for professional development put more burden on teachers and diminished the collaborative culture by establishing top-down management that positioned a head teacher as an evaluator. In Korea, teachers perceived the training that lower-skilled teachers received following a teacher evaluation as a punishment rather than constructive support. These two countries' cases highlight the importance of school climate and organizational context.

Unlike the other countries, in Finland, teacher evaluation was viewed as a teacher empowerment tool in the national-level policy initiatives. Finnish policymakers implemented teacher evaluation reform with a focus on professional growth beginning in 2000 and removed the state's inspection visit and rigid teacher evaluation schedules. Without a national framework or inspection, the municipality and principals conducted teacher evaluations. Principals were responsible for supporting the individual teacher in

deciding what types of professional development or training need to be followed based on their performance and development plans. In the 2013 TALIS survey, about three-quarters of principals responded that their teachers had been formally evaluated, which was relatively low compared to the other countries in which almost all the principals reported that formal teacher evaluation was conducted.

It is worth noting that in Finland, even though the country explicitly emphasized the purpose of professional growth, about half of the teachers reported that teacher appraisal and feedback had little impact upon the way teachers teach in the classroom in the TALIS survey. Compared to the teachers in the other three countries, the Finnish teachers reported the lowest incidence of professional growth outcomes. Like the other countries, there were inconsistencies between the principal and teacher survey results on the professional growth outcomes questionnaires, in which the principals viewed the professional development outcomes more positively than the teachers. Decentralizing the teacher evaluation system did not apparently have the intended effect of teacher empowerment and improving teacher quality but was rather seen by most teachers as an administrative requirement that had limited effects on teaching practice.

Finland was not, however, alone in this regard. The review of national policy reforms indicated that global policy debates around a “teacher quality” problem influenced the development of teacher evaluation policy reforms in all four countries. Noticeably, in the national policy narratives on teacher evaluation reform, policy actors referenced the discussion led by the OECD and its PISA rankings. However, the ways that these four countries reacted to the global dynamics were different. One focused on

increasing teacher accountability through teacher evaluation, another emphasized the role of teacher evaluation as an empowerment tool, and the other two attempted to achieve both. In addition, as seen in the policy analysis of Finland, South Korea, and Japan, efforts to improve teacher accountability did not necessarily result in linking teacher evaluation outcomes with high-stakes consequences, such as dismissal or the non-renewal of contracts. These non-U.S. countries implemented policy initiatives to uphold their teacher accountability while ensuring the job security of a teaching profession and thus a stable teaching workforce.

However, regardless of which approach each country took, a significant portion of teachers perceived that teacher evaluation and feedback were done to fulfill administrative requirements. Furthermore, the percentage of teachers that agreed or strongly agreed that the teacher evaluation and feedback had little impact on their teaching ranged between one-third and one half in each of the countries. These teacher survey results questioned the effectiveness of teacher evaluation reforms in practice. Even though each country devoted multiple years of effort to improve teaching and learning through teacher evaluation policy, it seems that the impact of those efforts in the classroom was limited. This calls for the need to more closely investigate the teacher evaluation policy process.

Reflection on Factors Affecting Successful Teacher Evaluation Implementation

While scholars have made important discoveries regarding both teacher policies and teachers' work within school organizations, limited research has linked these two bodies of literature. The implementation process is critical to the success of any policy.

However, it is difficult and complicated since a policy is often mediated by people in a targeted organization in various ways, which have been variously referred to as "slippage," "mutual adaptation," and "continuous improvement" (Ingram et al., 2004; Kraak, 2001; McLaughlin, 1990). Focusing on the actors in organizations is particularly important when studying teacher policies, as the goal is often not to change the structures but the people within them.

Using the 2013 TALIS teacher survey, the organizational factors that were associated with the positive impact of teacher evaluation based on teachers' perceptions were analyzed. Although the descriptive statistics indicated that a large portion of teachers did not perceive that the teacher evaluation and feedback that they received were effective in enhancing their teaching practices, the results of a structural equation model analysis showed that teacher evaluations had a positive impact in all the analyzed countries *under certain contextual organizational conditions*. First, the ways in which the teacher evaluation was implemented was an important factor. When teachers perceived that the outcomes of the teacher evaluation and feedback were linked to professional growth and support, they were more likely to perceive that the feedback they received had a positive impact on their teaching capacities. Second, the school climate in which the teacher evaluation was implemented was also critical. A school climate characterized by shared responsibility had meaningful relationships with teacher evaluation outcomes and positive relationships with professional capacity. More details on these key factors are discussed in the following sections.

Importance of Using Outcomes for Professional Growth

In the last two decades, policies have centered on teacher evaluation aimed at enhancing teaching practices. In many countries, particularly in the U.S. and England, teacher evaluation reform was implemented as a policy lever to uphold teacher accountability for student achievement. The singular focus connecting teacher evaluation, accountability, and tested student achievement as a means of achieving a high-quality teaching force has shifted due to growing concerns around the unintended consequences of high-stakes evaluation and the acknowledgment that accountability measures alone did not guarantee the improvement of teaching practices. This study, thus, investigated the professional growth outcomes of teacher evaluation and its relationships with teacher professional capacity.

The analysis showed that teachers in the four examined countries were more likely to perceive that teacher evaluation was positively associated with their professional capacities when it was coupled with explicit professional growth opportunities, which included actions, such as establishing a development/training plan for teachers to improve their work, providing feedback based on a thorough assessment of their teaching, discussions of ways to remedy any weaknesses in teaching, and the assignment of a mentor to help the teacher improve their teaching. All these measures highlight the educative purpose of teacher evaluation as part of an individually designed improvement process.

This result confirms recent scholarly work that pointed out the importance of "learning" as part of teacher evaluation (Darling-Hammond, 2013; Taylor & Tyler,

2012a). Teacher evaluation can influence teaching practices positively when it induces teacher learning through various outcomes. Advancing this notion of research, the findings of this study also shed light on how teacher evaluation can achieve the purpose of professional growth. It asserts that the outcomes of a teacher evaluation policy should support teacher professional growth, such as mentor assignment or training plans, to provide ample professional learning opportunities and considerable information. Even when it was accompanied by policies that used job-related sanctions and rewards (as it was in three of the four countries), the positive effects of an explicit professional development focus seem clear.

The Role of Shared Responsibility

While it is critical to link teacher evaluation and feedback with outcomes that support teacher professional growth, it does not guarantee the successful implementation of teacher evaluation. Lack of consideration of the organizational contexts in which teacher evaluation is implemented could result in a failure to achieve the goal of professional growth and yield unintended consequences. For instance, in Korea, training provided for low performers, which was intended to develop the teaching skills and competencies of teachers, was viewed as a punishment in some contexts because it was implemented as a requirement rather than as professional support and collaboration.

The findings from this study confirmed the significance of an organizational climate characterized by shared responsibility, showing a consistent positive relationship between shared responsibility and teacher evaluation and feedback variables in all four countries. Although the data in this study were cross-sectional, this suggests that shared

responsibility may have influenced the ways in which teacher evaluation was implemented, specifically in how outcomes would be used. The likelihood of teacher evaluation being followed by professional growth outcomes, such as training plans or mentor assignments, increases when a school developed a climate characterized by mutual support and collaboration, and staff members, parents, and students had opportunities to participate in the decision-making process on school issues.

The implications of the findings are also deepened by the positive association among shared responsibility, teacher evaluation policy, and increases in professional capacities. In previous research on teacher responses to high-stakes teacher evaluation, teachers often expressed increased stress as well as a disconnect with learning opportunities. In contrast, this study showed that school climate, as indicated by shared responsibility, contributed to perceptions of improved teaching practices, teacher motivation/confidence/job satisfaction, and teacher leadership. The results indicated that the organizational contexts of shared responsibility would be integral to amplify teacher learning opportunities and professional growth through teacher evaluation.

The findings are consistent with the evidence from numerous studies that have shown the importance of a collaborative culture and shared responsibility for teacher learning and professional development that were reviewed in Chapter 2. They also clearly showed that teacher evaluation should be implemented with a multidimensional approach to improve the core of teaching.

Teacher Evaluation and Growing Professional Capacity

This study examined three indicators of professional capacity, each of which has been associated with students' cognitive learning in other studies: (1) teaching practices, (2) teacher motivation and job satisfaction, and (3) teacher leadership. The professional capacity variable, which encompassed all three of these aspects of teaching, was significantly positively associated with other factors within the analytical model. These findings suggest that teacher evaluation policy should address a more comprehensive and wider range of professional capacities than has been previously discussed.

Much of the existing research on teacher evaluation has spotlighted its influence over teaching practices and student learning, which is essential. By strengthening the reflective practices through teacher evaluation and offering professional development support, teachers can reasonably improve classroom teaching practices and address issues in teaching and learning. At the same time, this approach may be limited in illuminating the comprehensive utility of teacher evaluation tools in schools. This study expands this notion by presenting that teacher evaluation could have more holistic effects not only on teachers' performance but also on their professional identity or career.

Teacher motivation/confidence/job satisfaction, as well as teacher leadership, have a profound impact on teacher retention and school improvement. The research on teacher retention has repeatedly shown the importance of teacher confidence, the sense of job satisfaction, and motivation toward continuous improvement (Billingsley, 2004; Boyd et al., 2011; Elfers, Plecki, & Knapp, 2006). The research on the National Schools

and Staffing Survey indicated that the main factors that drove teacher turnover were job dissatisfaction and teachers pursuing other jobs (Ingersoll, 2001).

Moreover, factors such as teacher learning, teacher leadership, and teacher retention are closely connected to other aspects of school improvement. For example, research has highlighted the role of collective teacher learning in promoting and sustaining teachers' engagement in intellectual and professional pursuits, enhancing teaching practice and student academic performance, retaining effective teachers, creating a cohesive and democratic culture among school staff, and providing venues for sharing norms and values for school reform (Coburn & Russell, 2008; Grossman, Wineburg, & Woolworth, 2001; Louis, Marks, & Kruse, 1996; Mulford, 2007; Stoll & Louis, 2007). Regarding teacher leadership, research on school improvement and reform has underscored the role of teacher leadership in contributing to school improvement and better student learning (Little, 2003; Muijs & Harris, 2003; Murphy, 2005; Smylie, Conley, & Marks, 2002; York-Barr & Duke, 2004). In sum, better evaluation policies and implementation at the school level could enhance the larger school improvement agenda, which is key to ensuring the stability and sustainability of school improvement efforts (Barnes, Crowe, & Schaefer, 2007; Billingsley, 2004; Carver-Thomas & Darling-Hammond, 2017; Guin, 2004).

Conclusion

There is no global policy consensus on what constitutes a good teacher evaluation policy nor on how it could be used to enhance teachers' work-life and their capacities. Teacher evaluation is widely used in every country, and almost all school leaders in this

study reported that their teachers have been evaluated. However, teachers perceived that it has not been effectively used: a significant portion of teachers in all countries reported that teacher evaluation in their schools was largely done to fulfill administrative requirements. Teacher evaluation has been problematic because conventional teacher evaluation provided scarce information for professional growth (Taylor & Tyler, 2012), and in teacher evaluation reform, the results were used to blame those who were marked with signs of ineffectiveness. This approach has limited the potential of teacher evaluation for professional growth. If teacher evaluation can be used primarily as one of the tools for professional capacity building, its scalable impact could be substantial.

Responding to global dynamics, the countries examined in this study wrestled with the teacher quality issue through national policy implementation that leveraged teacher accountability, professional development, or both. However, regardless of which direction they chose, the results suggest that their policy aspirations fell short. Thus, it is notable that teachers perceived that teacher evaluation was positively associated with teacher perceived professional capacities when it was coupled with actions to support their professional growth measures and shared responsibility at the local level. When these organizational conditions were met, the effects of teacher evaluation and feedback systems were broadly associated with teaching performance and teacher motivation/confidence/job satisfaction. This implies a more comprehensive impact of teacher evaluation on teaching and school improvements in which local implementation was attentive to teachers' needs.

Two organizational elements, professional growth outcomes and shared responsibility, were key factors in enhancing professional capacity from a teacher's perspective. The proposed analytical framework highlights the importance of both elements, suggesting that developing only one may not be sufficient in facilitating good teaching. Thus, I argue that the purpose of teacher evaluation should be reconceptualized as an educative tool in consideration of local contexts, providing ample feedback and opportunities for reflection that stimulate self-directed improvement. The literature review showed that teacher evaluation should be implemented with the balanced purposes of accountability and professional growth while moving away from the individualistic approach. To balance these two goals, the educative aspect of teacher evaluation, which has long been neglected, should be considered vital when implementing teacher evaluation.

Conducting teacher evaluation with good measures is important. However, that is not the end. The policy implementation of teacher evaluation, particularly how teacher evaluation outcomes are used for professional growth and to achieve its goals, could also significantly impact the professional capacity perceived by teachers. While policy implementation matters, the implementation should be considered in the context of shared responsibility and collaboration. This means principals and others who are involved in teacher evaluation need to be prepared to understand its purposes, as well as its procedures, and to give them the capacity to evaluate with supportive improvement in mind. Thus, its purpose should be reconceptualized, and its policy implementation contexts should be carefully considered. By doing so, teacher evaluation could yield a

more comprehensive impact on teaching and learning and on school improvement. This study suggests that teacher evaluation provides fertile ground for teachers to grow their professional capacity when thoughtfully designed and carefully implemented.

Theoretical Implications for Scholarship

This study advanced existing research by developing a theoretical framework based on the synthesis of work in three disciplines. Economics and human resources theory provided a meaningful perspective of teacher evaluation by critically reviewing teacher evaluation measures and broadening the goals of teacher evaluation to societal values. Teacher education scholars, who have investigated the core of teaching practices and performance, provided meaningful findings on the components of good teaching and the significance of teacher learning. Scholars studying organizational theory, which has focused on the organizational dynamics of the policy implementation process, took multi-dimensional approaches in the examination of effective teacher evaluation policy that would enhance teaching and learning in classrooms as well as aid in school-wide improvements (see Figure 6.1).

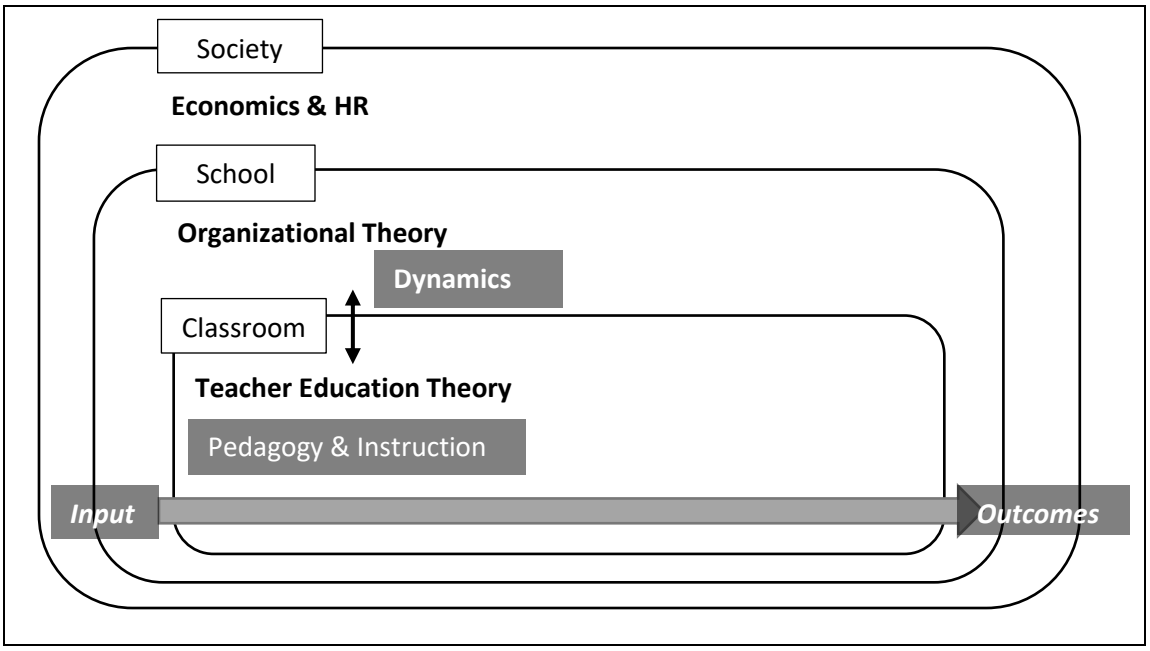


Figure 6.1 ***Different Theoretical Approaches to Teacher Evaluation Policy***

Often, their different approaches and arguments brought tension and contentious debates in the research circle among scholars, but this study has attempted to draw from all three, while centering the specific research questions within the organizational theory. This study aimed to investigate the associations among these three elements, particularly the implementation of teacher evaluation outcomes (drawing from both economics and teaching/teacher education), the school climate of shared responsibility (drawing from organizational theory), and their relationships with teachers and teaching. The results showed that these three elements were associated with one another. Figure 6.2 presents the synopsis of these findings based on the contributions of the three theoretical perspectives.

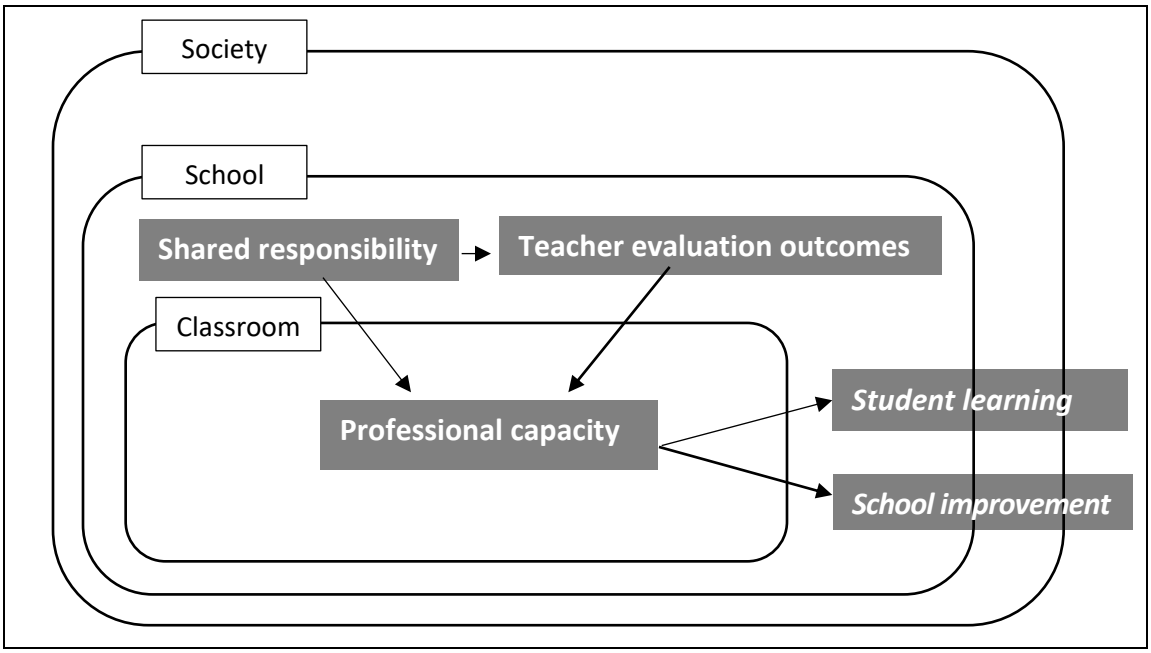


Figure 6.2 ***Synthesis of Theoretical Grounds and Findings of the Study***

Furthermore, by examining teacher evaluation in multiple countries, this study broadens the understanding of cross-national trends and the divergence of teacher evaluation policy, which are largely ignored in studies of single countries. All the countries examined in this study initiated and enacted national-level teacher evaluation reform influenced by global dynamics. In responding to the global dynamics in order to improve teacher quality, each country chose different solutions based on "within nation-specific teaching and policy environments" (Akiba, 2017, p. 153). However, this study also found that the significance of linking teacher evaluation to professional growth outcomes and developing shared responsibility were consistent in all four countries. The proposed analytical model showed a good fit in three out of the four countries, which included a European country, a North American country, and an Asian country. There was a marginal but consistent model fit in an additional Asian country. Although each of

the countries established distinct school culture and policy environments, the patterns of how the three latent variables associated with one another were similar across these countries. Accordingly, this study can contribute to generalizing an analytical model of teacher evaluation due to its application in multiple countries.

Implications for Policy Makers and School Leaders

The findings of this study have significant implications for policymakers and school leaders in the U.S. Often, the discussion of teacher evaluation policy centers on how teachers should be evaluated rather than what teacher evaluation outcomes should be used, what actions should be taken as a result of teacher evaluation, and what is to be accomplished. In contrast, this study sheds light on how policymakers and school leaders can successfully use teacher evaluation to support teachers' professional capacity.

First, this study affirms the recent trend in the U.S. and other countries to loosen the rein of individual teacher accountability at the national, state, and local levels and reshape teacher evaluation as a developmental feedback tool. The literature review made it clear that the individualistic approach to teacher accountability has not achieved the intended goal (i.e., better student learning) and has often yielded unintended consequences.

Moreover, this study provides meaningful insight into conditions that can facilitate good teaching and effective school improvement through teacher evaluation. This study expands the previous narrow perspective on evaluating teacher evaluation instruments (performance standards, psychometric validity and reliability, the methods of providing feedback, etc**.).** By situating evaluation in the organizational context of

teachers' work lives in a particular country and a particular school. It points to the differing perspectives of principals (the primary evaluator in all countries in this study) and teachers and links teachers' assessments of their experiences with evaluation to a key indicator of school climate—a shared responsibility. This points to the critical role of school leaders in developing effective teacher evaluation processes that are consistent with national policy but attentive to promoting teacher learning through feedback, supporting appropriate professional development plans, and providing mentoring opportunities. In addition, school leaders need to acknowledge that successful teacher evaluation should be grounded in the shared responsibility of actors in the organization. However, few principals currently in practice are prepared to carry out these expanded professional development roles. Thus, policymakers need to guide and support schools to create conditions that facilitate good teacher evaluation.

Regarding the utility of teacher evaluation, policymakers and school leaders should attend to the larger comprehensive impact of teacher evaluation. Teacher evaluation, unlike other professional measures, can be used as leverage for both teaching practices and the career advancement of teachers. Based on the results of teacher evaluation, school leaders may provide ample feedback and individualized support plans in a formative way. Also, the results could be used to match teachers to an appropriate role in school initiatives or different job responsibilities that are consistent with their strengths and provide opportunities for growth. Given the potential impact of teacher evaluation on various aspects of teaching, school leaders and policymakers may want to

strategize their teacher retention and school improvement efforts with in-depth considerations of teacher evaluation.

Implementation and financial feasibility are important considerations for school leaders and policymakers. In a practical sense, using an existing tool can be a cost-effective way to enhance teaching and the school environment. School leaders and policymakers have been involved in the work of teacher evaluation for many years—the idea is not new, although the approaches have been varied. Reshaping an existing tool to enhance the growth opportunities for individual teachers and school-wide improvement requires a new narrative to address the belief among teachers that it is an administrative task that has little connection to their daily work and, in some cases, is intended as a punishment. A dramatic reshaping that promotes meaningful teacher evaluation procedures and a developmental feedback loop could be costly but less expensive than developing new professional development or school improvement initiatives to be layered on top of existing evaluation requirements.

Future studies

There are several lines of further investigation that emerge from this effort. First, while the study provided meaningful insights on how to amplify professional growth opportunities through teacher evaluation, little was discussed on how to hold teachers accountable in a collaborative, shared responsibility culture. The study chose not to examine teachers' responses to high-stakes outcomes because an indicator was not formed that could be used across countries. However, it is known that a perceived conflict between high-stakes and developmental outcomes is important in some contexts

(Gilles, 2017). More in-depth contextual, culture-specific studies that examine how principals and teachers can work together to link evaluation with both accountability and professional growth are clearly needed.

Furthermore, this study aimed to expand a line of research that has been carried out primarily by North American scholars to include both policy and practitioner perspectives from systems that are politically and culturally very different. Future research that is more sensitive to the contexts of different educational systems could further enhance our understanding of the potential of teacher evaluation to support real school improvement.

Finally, this study focused on teacher evaluation outcomes primarily from the teachers' perspective (with some data from principals) and did not include the voices of students, parents, or other affected parties. Thus, although situated in an organizational rather than an individual framework, there are boundaries that need further exploration.

References

Akiba, M. (2017). Editor's Introduction: Understanding Cross-National Differences in Globalized Teacher Reforms: SAGE Publications Sage CA: Los Angeles, CA.

Alexander, N. A. (2012). Policy analysis for educational leaders: A step by step approach.

Alexander, N. A., Jang, S. T., & Kankane, S. (2017). The Performance Cycle: The Association between Student Achievement and State Policies Tying Together Teacher Performance, Student Achievement, and Accountability. *American Journal of Education, 123*(3), 413-446.

Ames, C. (1990). Motivation: What teachers need to know. *Teachers College Record, 91*(3), 409-421.

Aspinall, R. W. (2001). *Teachers' unions and the politics of education in Japan*: SUNY Press.

Ballou, D., & Springer, M. G. (2015). Using Student Test Scores to Measure Teacher Performance. *Educational Researcher, 44*(2), 77-86. doi:doi:10.3102/0013189X15574904

Barnes, G., Crowe, E., & Schaefer, B. (2007). The cost of teacher turnover in five school districts: A pilot study. *National Commission on Teaching and America's Future*.

Becker, A., Dumais, J., LaRoche, S., & Mirazchiyski, P. (2016). TALIS: User guide for the international database: Paris, France: Organisation for Economic Cooperation and Development (OECD ….

Beerens, D. R. (2000). *Evaluating Teachers for Professional Growth: Creating a Culture of Motivation and Learning*: ERIC.

Benchmarking, C. o. I. E. (2020). Top performing countries. Retrieved from https://ncee.org/what-we-do/center-on-international-education-benchmarking/top-performing-countries/

Berry, B., Darling-Hammond, L., Hirsch, E., Robinson, S., & Wise, A. (2006). No Child Left Behind and the „highly qualified" teacher: The promise and the possibilities. *Center for Quality Teaching, 2006*, 1-9.

Billingsley, B. S. (2004). Special education teacher retention and attrition: A critical analysis of the research literature. *The Journal of Special Education, 38*(1), 39-55.

Blase, J., & Blase, J. (1999). Principals' instructional leadership and teacher development: Teachers' perspectives. *Educational Administration Quarterly, 35*(3), 349-378.

Bolman, L. G., & Deal, T. E. (2013). *Reframing organizations* (5th Editon ed.).

Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). The influence of school administrators on teacher retention decisions. *American Educational Research Journal, 48*(2), 303-333.

Canrinus, E. T., Helms-Lorenz, M., Beijaard, D., Buitink, J., & Hofman, A. (2012). Self-efficacy, job satisfaction, motivation and commitment: Exploring the relationships between indicators of teachers' professional identity. *European journal of psychology of education, 27*(1), 115-132.

Carver-Thomas, D., & Darling-Hammond, L. (2017). Teacher turnover: Why it matters and what we can do about it: Palo Alto, CA: Learning Policy Institute.

Caughlan, S., & Jiang, H. (2014). Observation and teacher quality: Critical analysis of observational instruments in preservice teacher performance assessment. *Journal of teacher Education, 65*(5), 375-388.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling: a multidisciplinary journal, 14*(3), 464-504.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. Retrieved from

Choi, H. J., & Park, J.-H. (2016). An analysis of critical issues in Korean teacher evaluation systems. *Center for Educational Policy Studies Journal, 6*(2), 151-171.

Coburn, C. E., & Russell, J. L. (2008). District Policy and Teachers' Social Networks. *Educational Evaluation and Policy Analysis, 30*(3), 203-235. doi:10.3102/0162373708321829

Cochran-Smith, M., Piazza, P., & Power, C. (2013). The Politics of Accountability: Assessing Teacher Education in the United States. *The Educational Forum, 77*(1), 6-27. doi:10.1080/00131725.2013.739015

Cohen-Vogel, L. (2005). Federal role in teacher quality:"Redefinition" or policy alignment? *Educational Policy, 19*(1), 18-43.

Cohen, D. (2010). Teacher quality: An American educational dilemma. *Teacher assessment and the quest for teacher quality: A handbook*, 375-401.

Coolahan, J., Santiago, P., Phair, R., & Ninomiya, A. (2004). Country Note.

Danielson, C. (2008). *The handbook for enhancing professional practice: Using the framework for teaching in your school*: ASCD.

Danielson, C. (2013). *The framework for teaching: Evaluation instrument*: Danielson Group Princeton, NJ.

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*: Teachers College Press.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi delta kappan, 93*(6), 8-15.

Darling-Hammond, L., & Bransford, J. (2005). *Preparing teachers for a changing world: What teachers should learn and be able to do*: John Wiley & Sons.

Darling-Hammond, L., Burns, D., Campbell, C., Goodwin, A. L., Hammerness, K., Low, E.-L., . . . Zeichner, K. (2017). *Empowered Educators: How High-performing Systems Shape Teaching Quality Around the World*: John Wiley & Sons.

Davis, J., & Wilson, S. M. (2000). Principals' efforts to empower teachers: Effects on teacher motivation and job satisfaction and stress. *The clearing house, 73*(6), 349-353.

Demerath, P., Lynch, J., & Davidson, M. (2008). Dimensions of psychological capital in a US suburb and high school: Identities for neoliberal times. *Anthropology & Education Quarterly, 39*(3), 270-292.

Durksen, T. L., Klassen, R. M., & Daniels, L. M. (2017). Motivation and collaboration: The keys to a developmental framework for teachers' professional learning. *Teaching and Teacher Education, 67*, 53-66.

Eisenhardt, K. M. (1989). Agency Theory: An Assessment and Review. *The Academy of Management Review, 14*(1), 57-74. doi:10.2307/258191

Elfers, A. M., Plecki, M. L., & Knapp, M. S. (2006). Teacher Mobility: Looking More Closely at "The Movers" within a State System. *Peabody Journal of Education, 81*(3), 94-127.

Executive Office of the President. (2002). A Quality Teacher in Every Classroom: Improving Teacher Quality and Enhancing the Profession. Executive Office of the President, Washington, DC.

Firestone, W. A. (2014). Teacher Evaluation Policy and Conflicting Theories of Motivation. *Educational Researcher, 43*(2), 100-107. doi:10.3102/0013189x14521864

Ford, T. G., Urick, A., & Wilson, A. S. (2018). Exploring the effect of supportive teacher evaluation experiences on US teachers' job satisfaction. *education policy analysis archives, 26*, 59.

Fowler Jr, F. J. (2013). *Survey research methods*: Sage publications.

Geiser, C. (2012). *Data analysis with Mplus*: Guilford press.

Gilles, J. F. (2017). " It's Not a Gotcha": Interpreting Teacher Evaluation Policy in Rural School District. *The Rural Educator, 38*(2).

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Review of Economics and Statistics, 89*(1), 134-150.

Goldhaber, D., Harris, D. N., Loeb, S., McCaffrey, D. F., & Raudenbush, S. W. (2015). Carnegie Knowledge Network Concluding Recommendations. What We Know Series. *Carnegie Foundation for the Advancement of Teaching*.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make Room Value Added. *Educational Researcher, 44*(2), 96-104. doi:doi:10.3102/0013189X15575031

Goldstein, J. (2010). *Peer Review and Teacher Leadership: Linking Professionalism and Accountability. Series on School Reform*: ERIC.

Grossman, P., Wineburg, S., & Woolworth, S. (2001). Toward a Theory of Teacher Community. *The Teachers College Record, 103*, 942-1012.

Guin, K. (2004). Chronic teacher turnover in urban elementary schools. *education policy analysis archives, 12*, 42.

Hallgren, K., James-Burdumy, S., & Perez-Johnson, I. (2014). State Requirements for Teacher Evaluation Policies Promoted by Race to the Top. NCEE Evaluation Brief. NCEE 2014-4016. *National Center for Education Evaluation and Regional Assistance*.

Hanushek, E. A. (2003). The failure of input-based schooling policies. *The economic journal, 113*(485).

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. *Handbook of the Economics of Education*, *2*, 1051-1078.

Hargreaves, A., & Braun, H. (2013). *Data-driven improvement and accountability*. Retrieved from Boulder, CO: http://nepc.colorado.edu/publication/data-driven-improvement-accountability

Harris, A. (2003). Teacher leadership and school improvement. *Effective leadership for school improvement*, 72-83.

Harris, D. N. (2012). How Do Value-Added Indicators Compare to Other Measures of Teacher Effectiveness? What We Know Series: Value-Added Methods and Applications. Knowledge Brief 5. *Carnegie Foundation for the Advancement of Teaching*.

Harris, D. N., & Herrington, C. D. (2015). Editors' Introduction: The Use of Teacher Value-Added Measures in Schools. *Educational Researcher*, *44*(2), 71-76. doi:doi:10.3102/0013189X15576142

Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). *Applied survey data analysis*: CRC press.

Heinrich, C. J., & Marschke, G. (2010). Incentives and their dynamics in public sector performance management systems. *Journal of Policy Analysis and Management*, *29*(1), 183-208.

Hesterberg, T. (2011). Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(6), 497-526.

Hiebert, J., & Morris, A. K. (2012). Teaching, rather than teachers, as a path toward improving classroom instruction. *Journal of teacher Education, 63*(2), 92-102.

Honig, M. (2006). Complexity and policy implementation. *New directions in education policy implementation: Confronting complexity*, 1-25.

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal, 6*(1), 1-55.

Hull, J. (2013). Trends in teacher evaluation: How states are measuring teacher performance. *Center for Public Education Report*.

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal, 38*(3), 499-534.

Ingram, D., Louis, K. S., & Schroeder, R. G. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *Teachers College Record, 106*(6), 1258-1287.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher Perspectives on Evaluation Reform. *Educational Researcher, 44*(2), 105-116. doi:doi:10.3102/0013189X15575517

Judge, T. A., & Bono, J. E. (2001). Relationship of core self-evaluations traits—self-esteem, generalized self-efficacy, locus of control, and emotional stability—with job satisfaction and job performance: A meta-analysis. *Journal of applied Psychology, 86*(1), 80.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Paper presented at the Research Paper. MET Project. Bill & Melinda Gates Foundation.

Kang, N.-H. (2013). Teacher evaluation policy development in South Korea *Teacher reforms around the world: Implementations and outcomes*: Emerald Group Publishing Limited.

Katsuno, M. (2016). *Teacher evaluation policies and practices in Japan: How performativity works in schools*: Routledge.

Kennedy, M. M. (2008). Sorting out teacher quality. *Phi delta kappan, 90*(1), 59-63.

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher, 39*(8), 591-598.

Kim, J., & Youngs, P. (2016). Promoting instructional improvement or resistance? A comparative study of teachers' perceptions of teacher evaluation policy in Korea and the USA. *Compare: A Journal of Comparative and International Education, 46*(5), 723-744.

Kim, K., & Kim, E. (2012). The results of 2011 teacher evaluations. *Seoul: Korean Educational Development Institute*.

Kingdon, J. W. (1984). *Agendas, alternatives, and public policies*. Boston, MA: Brown & Co.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*: Guilford publications.

Knight, S. L., Lloyd, G. M., Arbaugh, F., Gamson, D., McDonald, S. P., Nolan Jr, J., & Whitney, A. E. (2015). Reconceptualizing teacher quality to inform preservice and inservice professional development: Sage Publications Sage CA: Los Angeles, CA.

Kraak, A. (2001). Policy Ambiguity and Slippage: Higher Education under the New State, 1994-2OOI. *Education in retrospect: Policy and implementation since 1990*, 85.

Labaree, D. F. (2012). *Someone has to fail: The zero-sum game of public schooling*: Harvard University Press.

Lipman, P. (2013). *The new political economy of urban education: Neoliberalism, race, and the right to the city*: Taylor & Francis.

Little, J. W. (2003). Constructions of teacher leadership in three periods of policy and reform activism. *School Leadership & Management, 23*(4), 401-419.

Lortie, D. C. (1975). *Schoolteacher: A sociological study*: JSTOR.

Louis, K. S., Febey, K., & Schroeder, R. (2005). State-mandated accountability in high schools: Teachers' interpretations of a new era. *Educational Evaluation and Policy Analysis, 27*(2), 177-204.

Louis, K. S., Marks, H. M., & Kruse, S. (1996). Teachers' Professional Community in Restructuring Schools. *American Educational Research Journal, 33*(4), 757-798.

McCaffrey, D. F., Lockwood, J., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability. Monograph*: ERIC.

McLaughlin, M. W. (1990). The Rand Change Agent Study Revisited: Macro Perspectives and Micro Realities. *Educational Researcher, 19*(9), 11-16. doi:10.3102/0013189x019009011

Morgan, G., Gregory, F., & Roach, C. (1997). *Images of organization*. Thousand Oaks: Sage Publication.

Moynihan, D. P. (2008). *The dynamics of performance management: Constructing information and reform*. Washingston, D.C.: Georgetown University Press.

Muijs, D., & Harris, A. (2003). Teacher Leadership—Improvement through Empowerment?: An Overview of the Literature. *Educational Management & Administration, 31*(4), 437-448. doi:10.1177/0263211x030314007

Mulford, B. (2007). Building social capital in professional learning communities: Importance, challenges and a way forward. *Professional learning communities: Divergence, depth and dilemmas*, 166-180.

Murphy, J. (2005). *Connecting teacher leadership and school improvement*: Corwin Press.

Murphy, J., Hallinger, P., & Heck, R. H. (2013). Leading via teacher evaluation: The case of the missing clothes? *Educational Researcher, 42*(6), 349-354.

Muthén, L., & Muthén, B. (2018). Mplus. *The comprehensive modelling program for applied researchers: user's guide, 5*.

National Board for Professional Teaching Standards. (2002). *What Teachers Should Know and be Able to Do*. Arlington, Va: National Board for Professional Teaching Standards.

National Education Association. (2015). Every Student Succeeds Act: Teacher Evaluations: National Education Association.

OECD. (2005). *Teachers matter: Attracting, developing and retaining effective teachers*: Organisation for Economic Co-operation and Development.

OECD. (2011). Finland: slow and steady reform for consistently high results. *Strong performers and successful reformers in education–lessons from PISA 2012 for the United States*, 117-136.

OECD. (2014). *TALIS 2013 Technical Report*. Retrieved from www.oecd.org/edu/school/TALIS-technical-report-2013.pdf

Paine, L., & Zeichner, K. (2012). The local and the global in reforming teaching and teacher education: University of Chicago Press Chicago, IL.

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123-141.

Pearson, L. C., & Moomaw, W. (2005). The relationship between teacher autonomy and stress, work satisfaction, empowerment, and professionalism. *Educational research quarterly*, *29*(1), 38-54.

Pennington, K., & Mead, S. (2016). For Good Measure?

Plecki, M., & Loeb, H. (2004). Lessons for policy design and implementation: Examining state and federal efforts to improve teacher quality. *Yearbook of the National Society for the Study of Education*, *103*(1), 348-389.

Plecki, M. L., Elfers, A. M., & Yeh, T. L. (2015). *Washington's Teacher and Principal Evaluation System: Efforts to Support Professional Development*. Retrieved from Seattle, WA:

Printy, S. M., Marks, H. M., & Bowers, A. J. (2010). Integrated leadership: How principals and teachers share transformational and instructional influence. *Jsl Vol 19-N5, 19*, 504.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review, 41*, 71-90.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*(2), 247-252.

Rockoff, J. E., & Speroni, C. (2011). Subjective and objective evaluations of teacher effectiveness: Evidence from New York City. *Labour Economics, 18*(5), 687-696.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics, 125*(1), 175-214.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31-57.

Sahlberg, P. (2011a). *Finnish lessons*: Teachers College Press.

Sahlberg, P. (2011b). The Professional Educator: Lessons from Finland. *American educator, 35*(2), 34-38.

Sato, M. (2014). What is the underlying conception of teaching of the edTPA? *Journal of teacher Education, 65*(5), 421-434.

Sato, M., Wei, R. C., & Darling-Hammond, L. (2008). Improving teachers' assessment practices through professional development: The case of National Board Certification. *American Educational Research Journal, 45*(3), 669-700.

Sawchuk, S. (2016). ESSA loosens reins on teacher evaluations, qualifications. *Education Week, 35*(15), 14-15.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online, 8*(2), 23-74.

Seo, K. (2012). Lessons from Korea. *Educational Leadership, 70*(3), 75-78.

Shafritz, J. M., Ott, J. S., & Jang, Y. S. (2011). *Classics of organization theory* (7th Ed ed.). Boston: Cengage Learning.

Smylie, M. A. (2014). Teacher Evaluation and the Problem of Professional Development. *Mid-Western Educational Researcher, 26*(2).

Smylie, M. A., Conley, S., & Marks, H. M. (2002). Exploring new approaches to teacher leadership for school improvement. *Yearbook of the National Society for the Study of Education, 101*(1), 162-188.

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*: Sage.

Spillane, J. P., Hopkins, M., Sweet, T. M., & Shirrell, M. (2017). The Social Side of Capability: Supporting Classroom Instruction and Enabling Its Improvement *Teaching in Context: The social side of education reform*. Cambridge, Massachusetts: Harvard Education Press.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy Implementation and Cognition: Reframing and Refocusing Implementation Research. *Review of Educational Research, 72*(3), 387-431. doi:10.3102/00346543072003387

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate behavioral research, 25*(2), 173-180.

Stoll, L., & Louis, K. S. (2007). *Professional learning communities: Divergence, depth and dilemmas*: McGraw-Hill Education (UK).

Strizek, G. A., Tourkin, S., & Erberber, E. (2014). Teaching and Learning International Survey (TALIS) 2013: US Technical Report. NCES 2015-010. *National Center for Education Statistics*.

Sun, M., Mutcheson, R. B., & Kim, J. (2015). Teachers' Use of Evaluation for Instructional Improvement and School Supports for This Use. *Making the Most of Multiple Measures: The Impacts and Challenges of Implementing Rigorous Teacher Evaluation Systems*.

Superfine, B. M., Gottlieb, J. J., & Smylie, M. A. (2012). The expanding federal role in teacher workforce policy. *Educational Policy, 26*(1), 58-78.

Sutton, M., & Levinson, B. A. (2001). *Policy as practice: Toward a comparative sociocultural analysis of educational policy* (Vol. 1): Greenwood Publishing Group.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5): Pearson Boston, MA.

Tarhan, H., Karaman, A., Lauri, K., & Aerila, J.-A. (2019). understanding teacher evaluation in Finland: A professional development framework. *Australian Journal of Teacher Education (Online), 44*(4), 33.

Taylor, E. S., & Tyler, J. H. (2012a). Can Teacher Evaluation Improve Teaching? *Education Next, 12*(4).

Taylor, E. S., & Tyler, J. H. (2012b). The effect of evaluation on teacher performance. *American Economic Review, 102*(7), 3628-3651.

Thoonen, E. E., Sleegers, P. J., Oort, F. J., Peetsma, T. T., & Geijsel, F. P. (2011). How to improve teaching practices: The role of teacher motivation, organizational factors, and leadership practices. *Educational Administration Quarterly, 47*(3), 496-536.

Waller, W. (1932). *The sociology of teaching*. New York: Russell and Russell.

Webb*, R., Vulliamy, G., Hämäläinen, S., Sarja, A., Kimonen, E., & Nevalainen, R. (2004). A comparative analysis of primary teacher professionalism in England and Finland. *Comparative education, 40*(1), 83-107.

Yoo, J. (2018). Evaluating the new teacher evaluation system in South Korea: Case studies of successful implementation, adaptation, and transformation of mandated policy. *Policy Futures in Education, 16*(3), 277-290.

York-Barr, J., & Duke, K. (2004). What Do We Know About Teacher Leadership? Findings From Two Decades of Scholarship. *Review of Educational Research, 74*(3), 255-316. doi:10.3102/00346543074003255

Appendix

Appendix A. Reliability of scaled survey items

Scale	Item	Alpha			
		Fin	Jap	Kor	U.S.
Perceived positive impact – Motivation, Confidence, Satisfaction	Confidence as a teacher Job satisfaction Motivation	.903	.874	.893	.911
Perceived positive impact- teaching practices	Classroom management practices Knowledge and understanding of your main subject field(s) Teaching practices Methods for teaching students with special needs Use of student assessments to improve student learning	.865	.786	.921	.896
Perceived positive impact on teacher leadership	Public recognition from the principal and/or your colleagues Role in school development initiatives Career advancement likelihood Amount of professional development Job responsibilities at this school	.820	.823	.895	.883
Teacher eval. outcomes- professional growth	Development/training plan Thorough Feedback Weakness remedy measures Mentor assignment	.696	.705	.795	.799
Shared responsibility	Staff participation in school decisions Parents/guardians participation in school decisions Students participation in school decisions Culture of shared responsibility for school issues Collaborative culture with mutual support	.815	.786	.880	.891

Appendix B. Definition of Terms

When developing the conceptual and analytical framework of this study, this paper defines the key terms of the framework as follows:

- **Professional capacity:** This paper defines that the desirable outcome of teacher evaluation policy is growing teacher professional capacity. In this paper, professional capacity is the term that encompasses three aspects of teaching (1) teaching practices in classrooms (2) teacher motivation, job satisfaction, and confidence, and (3) teacher leadership. This term conceptualizes teaching as an organizational and comprehensive construct.
- **Outcomes of teacher evaluation and feedback system:** Based on the recent research reviewed above, this paper argues that achieving dual purposes of accountability and professional growth are two main purposes of teacher evaluation. While the term "accountability" is consistently used across the literature, different terms are used to describe the professional growth aspect of teacher evaluation purposes, including improvements, professional learning, and professional development. This paper uses the term professional growth to emphasize both aspects of learning and improvements and to encompass the various aspects of professional capacity. This paper uses the definition in Plecki, Elfers and Yeh's report to define teacher evaluation for accountability and teacher evaluation for professional growth teacher evaluation systems as follows (2015, p.3-4):
    - **Evaluation for accountability:** Policymakers and school leaders "focus on using teacher evaluation to make decisions about hiring, firing, tenure or salary. This implies a high-stakes system of evaluation".
    - **Evaluation for professional growth:** Policymakers and school leaders use the teacher evaluation process to inform decisions about the kinds of professional learning and leadership opportunities "needed to help teachers and schools engage in continuous improvement".

- **Shared responsibility:** In this paper, the shared responsibility refers to the participatory and collaborative culture among local actors around teacher evaluation. In other words, this refers to a school climate that allows teachers, parents, and students to participate actively in the decision-making process in schools and promotes collaboration and mutual support within the school. According to the literature reviewed in this dissertation, a shared responsibility is key to achieving both goals of accountability and teacher professional growth.

Appendix C. Descriptive statistics of variables

Country	Construct	Mean	SD	Skewness	Kurtosis
Finland	Teacher eval. outcomes-professional growth	2.21	0.796	-0.046	-0.814
		1.87	0.694	0.335	-0.352
		2.64	0.711	-0.591	0.175
		1.76	0.759	0.651	-0.316
	Shared responsibility	2.81	0.661	-0.631	0.849
		2.61	0.640	-0.300	-0.046
		2.68	0.613	-0.392	0.173
		2.88	0.603	-0.615	1.373
		2.95	0.654	-0.467	0.759
	Perceived positive impact - Confidence, Job satisfaction and motivation	2.74	0.813	-0.215	-0.451
		2.68	0.843	-0.149	-0.586
		2.71	0.850	-0.176	-0.604
	Perceived positive impact-teaching practices	2.20	0.758	0.196	-0.326
		2.16	0.810	0.231	-0.521
		2.31	0.743	0.172	-0.239
		2.13	0.779	0.242	-0.420
		2.17	0.791	0.262	-0.382
	Perceived positive impact on teacher leadership	2.59	0.832	-0.121	-0.537
		2.14	0.854	0.278	-0.648
		1.65	0.798	1.057	0.428
		2.01	0.811	0.343	-0.619
		2.15	0.857	0.228	-0.731
Japan	Teacher eval. outcomes-professional growth	2.39	0.700	-0.174	-0.371
		2.22	0.638	-0.014	-0.290
		2.70	0.676	-0.838	0.695
		2.12	0.759	0.084	-0.663
	Shared responsibility	2.79	0.620	-0.756	1.185
		2.62	0.588	-0.536	-0.003
		2.46	0.650	-0.257	-0.303
		2.68	0.636	-0.630	0.453
		2.92	0.615	-0.566	1.293
	Perceived positive impact - Confidence, Job satisfaction and motivation	3.10	0.707	-0.618	0.600
		2.99	0.795	-0.554	0.000
		3.07	0.764	-0.612	0.183
	Perceived positive impact-teaching practices	2.80	0.848	-0.553	-0.160
		3.11	0.696	-0.632	0.738
		3.17	0.663	-0.560	0.662
		2.68	0.804	-0.308	-0.312
		2.86	0.711	-0.513	0.451
	Perceived positive impact on teacher leadership	3.07	0.760	-0.708	0.503
		2.66	0.779	-0.388	-0.174
		2.15	0.869	0.289	-0.671
		2.31	0.798	0.001	-0.577
		2.79	0.781	-0.511	0.068
Korea	Teacher eval. outcomes-professional growth	2.71	0.658	-0.643	0.555
		2.47	0.696	-0.158	-0.263
		2.78	0.647	-0.817	1.147
		2.38	0.744	-0.187	-0.488
	Shared responsibility	2.63	0.729	-0.549	0.056
		2.88	0.590	-0.712	1.670
		2.71	0.658	-0.613	0.507
		2.73	0.661	-0.701	0.688
		2.76	0.674	-0.718	0.806

	Perceived positive impact - Confidence, Job satisfaction and motivation	2.76	0.827	-0.311	-0.399
		2.50	0.874	-0.137	-0.690
		2.61	0.830	-0.157	-0.514
	Perceived positive impact-teaching practices	2.60	0.784	-0.207	-0.348
		2.68	0.758	-0.256	-0.207
		2.73	0.750	-0.216	-0.221
		2.67	0.767	-0.208	-0.280
		2.60	0.786	-0.245	-0.337
	Perceived positive impact on teacher leadership	2.60	0.749	-0.317	-0.185
		2.47	0.811	-0.245	-0.545
		2.13	0.917	0.207	-1.019
		2.55	0.799	-0.177	-0.431
		2.74	0.790	-0.279	-0.290
United States	Teacher eval. outcomes-professional growth	2.50	0.783	-0.407	-0.413
		2.48	0.801	-0.243	-0.497
		2.74	0.727	-0.625	0.391
		2.47	0.836	-0.222	-0.604
	Shared responsibility	2.68	0.771	-0.555	0.033
		2.80	0.702	-0.575	0.532
		2.56	0.719	-0.253	-0.196
		2.68	0.747	-0.397	-0.035
		2.74	0.741	-0.438	0.089
	Perceived positive impact - Confidence, Job satisfaction and motivation	2.71	1.011	-0.286	-1.011
		2.44	1.033	0.028	-1.157
		2.52	1.049	-0.074	-1.188
	Perceived positive impact-teaching practices	2.31	0.947	0.181	-0.901
		2.12	1.025	0.414	-1.035
		2.55	0.950	-0.123	-0.904
		2.13	0.964	0.372	-0.899
		2.44	0.976	-0.008	-1.008
	Perceived positive impact on teacher leadership	2.28	1.037	0.212	-1.153
		2.18	1.019	0.269	-1.142
		1.84	0.998	0.821	-0.591
		2.03	0.954	0.480	-0.833
		2.20	1.041	0.306	-1.133

Appendix D. Model fit of competing CFA models

Country	CFI	TLI	RMSEA	SRMR
U.S.	.837	.819	.070	.051
Finland	.799	.777	.068	.056
Korean	.895	.883	.065	.038
Japan	.794	.771	.078	.054

Appendix E. Factor loadings for latent variables

U.S.

Scale	Variable	Item	Factor loading
Teacher eval. outcomes-professional growth	TT2G31D	Development/training plan	.713
	TT2G31E	Thorough Feedback	.736
	TT2G31G	Weakness remedy measures	.745
	TT2G31H	Mentor assignment	.630
Shared responsibility	TT2G44A	Staff participation in school decisions	.796
	TT2G44B	Parents/guardians participation in school decisions	.701
	TT2G44C	Students participation in school decisions	.716
	TT2G44D	Culture of shared responsibility for school issues	.888
	TT2G44E	Collaborative culture with mutual support	.828
Perceived positive impact – Motivation, Confidence, Satisfaction	TT2G30F	Confidence as a teacher	.758
	TT2G30M	Job satisfaction	.946
	TT2G30N	Motivation	.948
Perceived positive impact- teaching practices	TT2G30H	Classroom management practices	.817
	TT2G30I	Knowledge and understanding of your main subject field(s)	.823
	TT2G30J	Teaching practices	.886
	TT2G30K	Methods for teaching students with special needs	.778
	TT2G30L	Use of student assessments to improve student learning	.780
Perceived positive impact on teacher leadership	TT2G30A	Public recognition from the principal and/or your colleagues	.740
	TT2G30B	Role in school development initiatives	.780
	TT2G30C	Career advancement likelihood	.736
	TT2G30D	Amount of professional development	.794
	TT2G30E	Job responsibilities at this school	.831

Finland

Scale	Variable	Item	Factor loading
Teacher eval. outcomes-professional growth	TT2G31D	Development/training plan	.646
	TT2G31E	Thorough Feedback	.669
	TT2G31G	Weakness remedy measures	.559
	TT2G31H	Mentor	.541
Shared responsibility	TT2G44A	Staff participation in school decisions	.776
	TT2G44B	Parents/guardians participation in school decisions	.671
	TT2G44C	Students participation in school decisions	.629
	TT2G44D	Culture of shared responsibility for school issues	.708
	TT2G44E	Collaborative culture with mutual support	.643
Perceived positive impact -Motivation, Confidence, Satisfaction	TT2G30F	Confidence as a teacher	.726
	TT2G30M	Job satisfaction	.941
	TT2G30N	Motivation	.950
Perceived positive impact- teaching practices	TT2G30H	Classroom management practices	.775
	TT2G30I	Knowledge and understanding of your main subject field(s)	.814
	TT2G30J	Teaching practices	.830
	TT2G30K	Methods for teaching students with special needs	.713
	TT2G30L	Use of student assessments to improve student learning	.738
Perceived positive impact on teacher leadership	T2G30A	Public recognition from the principal and/or your colleagues	.681
	TT2G30B	Role in school development initiatives	.675
	TT2G30C	Career advancement likelihood	.611
	TT2G30D	Amount of professional development	.720
	TT2G30E	Job responsibilities at this school	.735

Korea

Scale	Variable	Item	Factor loading
Teacher eval. outcomes-professional growth	TT2G31D	Development/training plan	.697
	TT2G31E	Thorough Feedback	.789
	TT2G31G	Weakness remedy measures	.671
	TT2G31H	Mentor assignment	.657
Shared responsibility	TT2G44A	Staff participation in school decisions	.785
	TT2G44B	Parents/guardians participation in school decisions	.640
	TT2G44C	Students participation in school decisions	.748
	TT2G44D	Culture of shared responsibility for school issues	.830
	TT2G44E	Collaborative culture with mutual support	.831
Perceived positive impact – Motivation, Confidence, Satisfaction	TT2G30F	Confidence as a teacher	.819
	TT2G30M	Job satisfaction	.869
	TT2G30N	Motivation	.908
Perceived positive impact- teaching practices	TT2G30H	Classroom management practices	.824
	TT2G30I	Knowledge and understanding of your main subject field(s)	.892
	TT2G30J	Teaching practices	.900
	TT2G30K	Methods for teaching students with special needs	.841
	TT2G30L	Use of student assessments to improve student learning	.830
Perceived positive impact on teacher leadership	TT2G30A	Public recognition from the principal and/or your colleagues	.798
	TT2G30B	Role in school development initiatives	.815
	TT2G30C	Career advancement likelihood	.704
	TT2G30D	Amount of professional development	.845
	TT2G30E	Job responsibilities at this school	.827

Japan

Scale	Variable	Item	Factor loading
Teacher eval. outcomes-professional growth	TT2G31D	Development/training plan	.700
	TT2G31E	Thorough Feedback	.741
	TT2G31G	Weakness remedy measures	.450
	TT2G31H	Mentor	.544
Shared responsibility	TT2G44A	Staff participation in school decisions	.696
	TT2G44B	Parents/guardians participation in school decisions	.577
	TT2G44C	Students participation in school decisions	.606
	TT2G44D	Culture of shared responsibility for school issues	.732
	TT2G44E	Collaborative culture with mutual support	.636
Perceived positive impact-Motivation, Confidence, Satisfaction	TT2G30F	Confidence as a teacher	.692
	TT2G30M	Job satisfaction	.920
	TT2G30N	Motivation	.917
Perceived positive impact- teaching practices	TT2G30H	Classroom management practices	.621
	TT2G30I	Knowledge and understanding of your main subject field(s)	.775
	TT2G30J	Teaching practices	.779
	TT2G30K	Methods for teaching students with special needs	.652
	TT2G30L	Use of student assessments to improve student learning	.779
Perceived positive impact on teacher leadership	T2G30A	Public recognition from the principal and/or your colleagues	.652
	TT2G30B	Role in school development initiatives	.704
	TT2G30C	Career advancement likelihood	.673
	TT2G30D	Amount of professional development	.719
	TT2G30E	Job responsibilities at this school	.735

Appendix F. SEM results by country

	U.S.		Finland		Korea		Japan	
	Coef.	p-value	Coef.	p-value	Coef.	p-value	Coef.	p-value
EP Direct	.444	< .001	.407	< .001	.438	< .001	.380	< .001
SP Direct	.243	< .001	.157	< .001	.174	< .001	.239	< .001
SE Direct	.551	< .001	.505	< .001	.519	< .001	.556	< .001
YP Direct	-.088	.006	-.020	.468	-.018	.369	-.061	.002
SEP Indirect	.244	< .001	.205	< .001	.227	< .001	.211	
SEP Total	.488	< .001	.362	< .001	.401	< .001	.450	
	Model fit		Model fit		Model fit		Model fit	
RMSEA	.040		.040		.062		.056	
CFI	.951		.939		.903		.893	
TLI	.944		.931		.891		.879	
SRMR	.038		.042		.091		.047	

Note. EFP = Evaluation and Feedback Outcomes for Professional Growth; MSC = Positive Impact on Motivation, Job Satisfaction, and Confidence; PC = Professional Capacity; SR = Shared Responsibility; TL = Positive Impact on Teacher Leadership; TP = Positive Impact on Teaching Practices; and YE = Years of Experience